PERFORMANCE OF GENOMIC PREDICTION FOR A SUGARCANE COMMERCIAL BREEDING PROGRAM

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Itaraju Junior Baracuhy Brum

August 2018

PERFORMANCE OF GENOMIC PREDICTION FOR A SUGARCANE COMMERCIAL BREEDING PROGRAM

Itaraju Junior Baracuhy Brum, Ph.D.

Cornell University 2018

Sugarcane is a clonally propagated crop of economic importance in tropical areas and is mostly used for production of sugar, ethanol, energy and animal feed. Cultivars are hybrids between two autopolyploid species, the domesticated "noble cane" *Saccharum officinarum* L. (2n=80) and the wild *Saccharum spontaneum* L. (2n=40-128). In this study genomic selection was evaluated as a tool to increase efficiency in the breeding program. A population of 1882 clones from two breeding cycles was genotyped by sequencing resulting in a filtered set of 55k SNPs, providing extensive genome coverage. This population was phenotyped for plot weight, Brix, fiber and sucrose content, with replicated measurements taken on first season crop and ratoon crop harvests. Broad-sense heritabilities ranged from 0.69 to 0.90. Genomic prediction accuracy was assessed with genomic best linear unbiased prediction models in two ways: for *clonal prediction* of the genotyped clones and for *parental prediction* of their respective progenitors. In *clonal prediction* accuracies ranged from 0.07 to 0.39 in cross validation within a breeding cycle, and 0.01 to 0.32 in predictions across cycles. In *parental prediction* accuracies varied from 0.14 to 0.17 for Brix, and from 0.20 to 0.26 for plot weight. We observed a strong genotype by year interaction effect leading to reduced accuracies when predicting across breeding cycles. The genomic predicted breeding value using progeny data, achieved similar accuracies as *clonal prediction*. These results could be taken into account

in the deployment of genomic selection for a sugarcane breeding program. We also investigated the use of high dosage information in the representation of SNP data from sugarcane. Association analysis and genomic prediction were performed using four fiber traits, for a *continuous* marker representation that can represent high dosage of alleles, and for a *discrete* representation, that is limited in distinguishing heterozygous from homozygous states. We observed an increase in the number of significant hits in association tests when using dosage coding. In genomic prediction, differences were small between *continuous* and *discrete* coding, but in most of the cases there was an advantage when using *continuous* coding.

BIOGRAPHICAL SKETCH

Itaraju Brum was born in Brazil on April 12, 1979, to parents Itaraju Pinto Brum and Uiara F. Baracuhy Brum. In 2004 Itaraju graduated on Computer Engineering at Universidade Estadual de Campinas (UNICAMP). And in 2007 he received a master degree in Molecular and Functional Biology from the same university. As a bioinformatist he was hired by CTC - Centro de Tecnologia Canavieira (Sugacane Technology Center) in 2007. With support from CTC, he and his wife, Sueli Mori Brum, moved to Ithaca in 2014, and he joined Mark Sorrell's lab pursuing the program in Plant Breeding and Genetics in Cornell University.

To my parents and wife.

"Knowledge alone does not produce wisdom. Transforming knowledge into wisdom requires input from the heart." (Daisaku Ikeda)

"The proud mission of those who have been able to receive education must be to serve, in seen and unseen ways, the lives of those who have not had this opportunity." (D.I.)

ACKNOWLEDGEMENTS

I would like to acknowledge all of those who contributed to this thesis and my program in Cornell University. First, I would like to thank Dr. Mark Sorrells, my special committee chair, for accepting me in his program, guidence and support in my research, and specially for inspiring me as a plant breeder. Also I would like to thank my special committee members, Dr. Jean-luc Jannink for his comments and suggestions on my research; and Dr. James Booth for this contributions to my thesis. I would like to thank students and post docs from Sorrells and Jannink lab, specially Dr. Deniz Akdemir, Dr. Uche Okeke, Dr. Nicholas Santantonio, whith whom I had the opportunity to share my research ideas and whom I got valuable suggestions. I would like to thank all the Sorrells Lab for the opportunity to work together on the wheat and barley fields, providing crucial experience for my development as plant breeder and scientist. My Ph.D program was sponsored by CTC (Centro de Tecnologia Canavieira, Piraciaba, Brazil), which also provided the field data, the molecular marker data and information from its breeding program that was used in this thesis. From CTC, I would like to thank specially Dr. Karine M. Oliveira, Dr. Francisco Claudio Lopes, Dr. Thiago R. Benatti who worked in stablishing the first Genomic Selection projects at CTC and helped to stablish the framework for the current research in this thesis, and worked on genotyping the sugarcane population used in the thesis. Also from CTC, I would like to thank Dr. Sabrina Chabregas who first sponsored my proposal of doing a Ph.D in Cornell, and also Dr. Michael Butterfield and William Burnquist who have been directly responsible for sponsoring my program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Sugarcane: the crop

Sugarcane is cultivated in tropical and subtropical regions across the world, the top producer countries being Brazil, India and China (fig. 1.1). Despite not being among the top 10 crops in cultivated area in the world (fig. 1.2), its high productivity makes it the first in total production. In 2016, world total sugarcane production was 1,890,661,751 tonnes and the second crop, maize production, was 1,060,107,470 tonnes (fig. 1.3) (FAO, 2018). Sugarcane uses mainly sucrose for energy storage, which is the component extracted for its most common uses: production of sugar and ethanol. Outside the sugar industry, sugarcane is also used for animal feed, and human consumption as juice, sweets and alcoholic beverages. Ethanol from sugarcane is also produced as biofuel in Brazil, considered an important form of reduction in greenhouse-gas emissions (Goldemberg et al., 2008). As an energy source, the sugar industry also utilizes the fiber, that remains from juice extraction, for electricity production.

Sugarcane related wild species are *Saccharum spontaneum* L., that has center of origin and diversity in India, and a broad distribution in tropical and subtropical regions, and *S. robustum* Brandes and Jewiet ex Grassl, with a center of diversity in New Guinea (Ming et al., 2010).

The cultivated species for sugar are *S. officinarum* L., from New Guinea, *S. barberi* Jeswiet, from India, and *S. sinense* Roxb. from China. *S. edule* Hassk., cultivated from New Guinea to Fiji, is used as a vegetable (Ming et al., 2010). *S. officinarum* is thought to have been derived from *S. robustum*, whereas *S. barberi* and *S. sinense* were derived from hybridization between *S. officinarum* and *S. spontaneum* as revealed from *in situ* hybridization results (D'Hont et al., 2002).

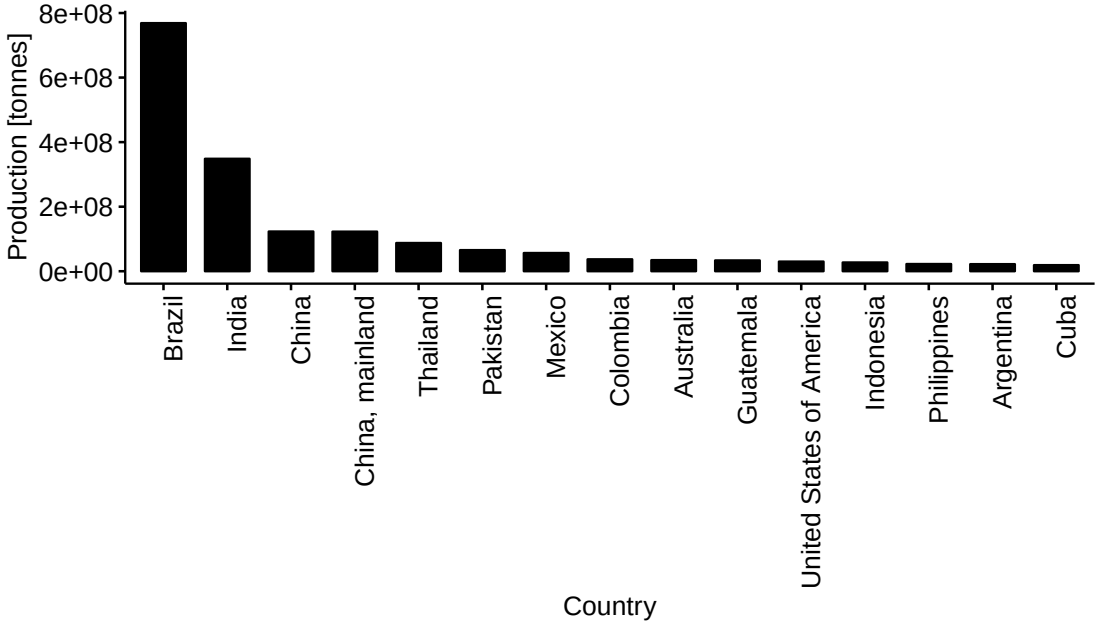

Figure 1.1: Sugarcane production (in total tonnes) per country in 2016 (last available data). Showing top 15 countries (FAO, 2018).

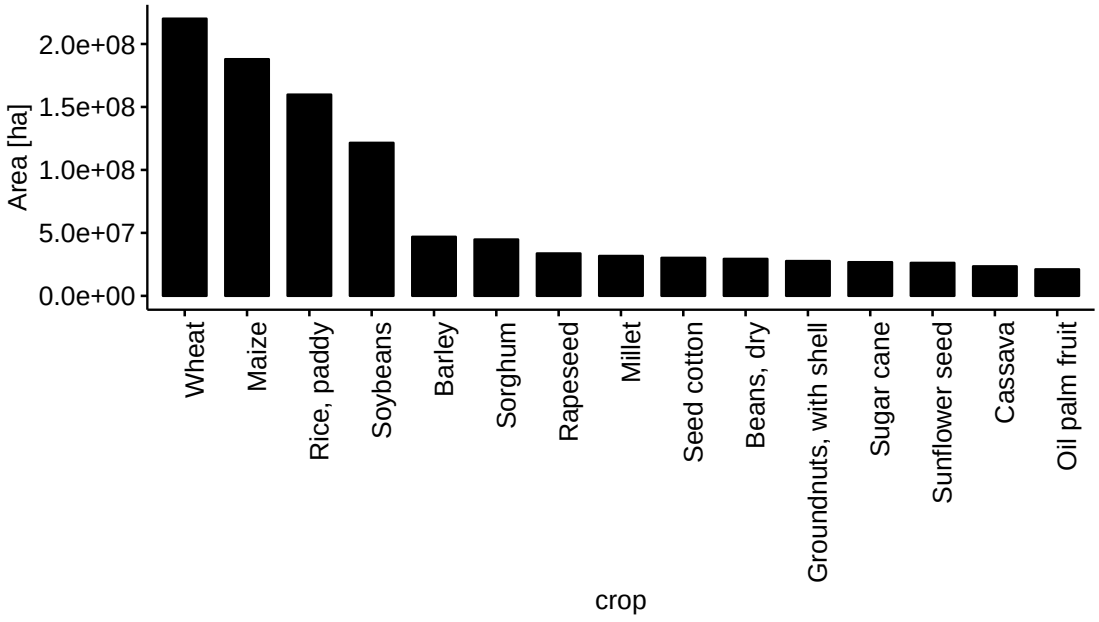

Figure 1.2: Planted area (in hectares) per crop in 2016 (last available data). Showing top 15 crops (FAO, 2018).

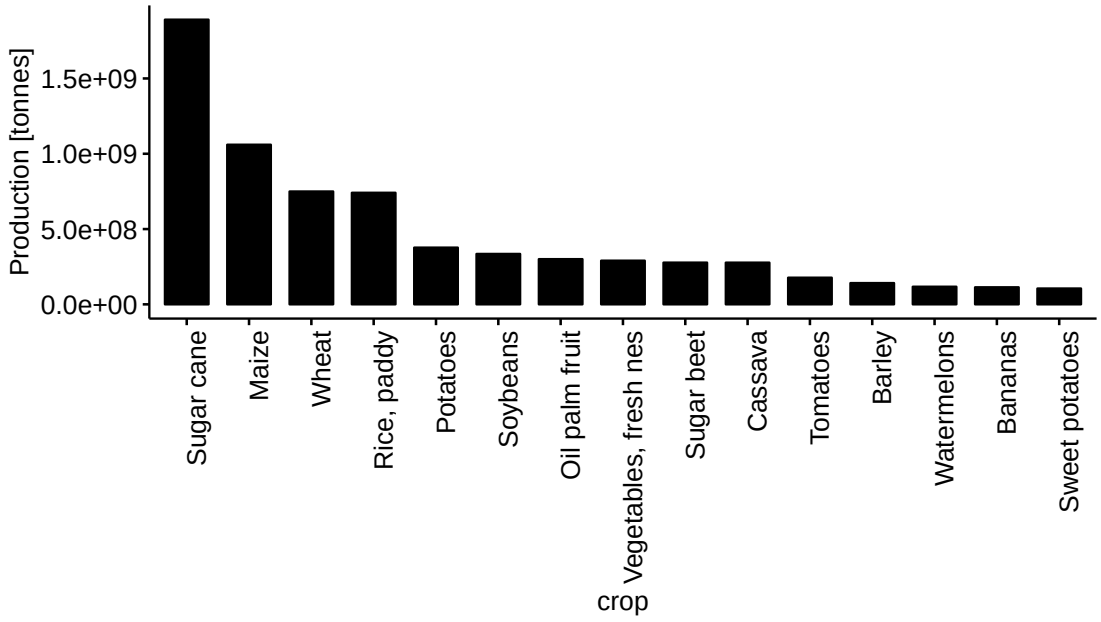

Figure 1.3: Production (in total tonnes) per crop in 2016 (last available data). Showing top 15 crops (FAO, 2018).

Irvine (1999) proposed that the genus Saccharum be divided into two species: *S. spontaneum* and *S. officinarum*.

The first breeding programs of sugarcane started in 1888, after two independent observations that it was possible to produce viable seed in crosses, in 1858 in Java, and 1859 in Barbados (Ming et al., 2010). Early breeding utilized *S. spontaneum*, for resistance to diseases and stresses, in crosses with *S. officinarum*, with hybrids being backcrossed to *S. officinarum* to retain sugar productivity. Those initial crosses were so successful that the crosses were not repeated even though they could have contributed to broadening the genetic basis (Jackson, 2005). Subsequent breeding mainly focused on crosses between the new hybrids.

Given that the original species are autopolyploids, so are the modern sug-

arcane cultivars, with chromosome numbers from 100 to 130 (Piperidis et al., 2010). Its huge genome size, the polyploidy and other features such as aneuploidy contribute to complexity in any investigation of sugarcane's molecular make up. In spite of that, recently sugarcane had its first genome reference sequence published (Garsmeur et al., 2018), covering a monoploid version of the polyploid chromosome set.

1.2 Why genomic selection in sugarcane?

Genomic selection was presented in the seminal paper by Meuwissen et al. (2001), where it was proposed to use the whole set of DNA derived molecular marker information to predict total genetic values or breeding values. The molecular markers should be widely distributed across the genome, and due to the consequent large number of features involved, the statistical models used had to perform feature selection of covariates to be included (exemplified in the least squares estimation with features selected by association analysis), or introduce some form of penalization in the model complexity, as presented in the methods BLUP, BayesA and BayesB (Meuwissen et al., 2001).

Reduction in genotyping costs was necessary for actual implementation of those ideas, with progress first in animal breeding and then in plants as well (Lorenz et al., 2011, Heslot et al. (2015)).

The long breeding cycles from sugarcane (more information is presented on chapter 3) would benefit from early and accurate selection of clones if provided by molecular makers (Jackson, 2005). Studies on association of traits and molecular markers have been successful only in few traits, with more limited success

in the case of complex traits (more on chapter 2). On the other hand, genomic selection has the potential to tackle complex traits, under the assumption that a large number of genomic regions with small effects affect the trait.

1.3 Thesis structure

The thesis was structured in four chapters, the current chapter and three written in the form of publications. Chapter 2 evaluates the use of genomic selection for sugarcane in order to predict clone performance. Chapter 3 evaluates genomic prediction in the context of selection of parents for a breeding program, and chapter 4 makes use of genomic prediction and association tests to confirm whether the encoding of dosage information in molecular marker data is useful for quantitative studies in sugarcane.

Genomic selection for sugarcane has been first described in the paper by Gouy et al. (2013), utilizing a collection of clones from different countries around the World. A prediction of total genetic values of the clones was also reported. This is similar to our analyses in chapter 2, but utilizing clones from a sugarcane commercial breeding program in Brazil. We also analyzed different traits, among them yield that is relevant for clone selection, and requires larger plots not available in the study by Gouy et al. (2013).

On chapter 3, we move on to a kind of analysis not yet available in the sugarcane literature, that analyzes the performance of genomic prediction for selection of parents. This provides another opportunity for application of genomic selection in a breeding program, its advantage in comparison to clone prediction is presented in the corresponding conclusion section from this chapter.

All chapters, 2, 3 and 4, make use of the same datasets of molecular markers and field phenotypic records. These datasets will be described in detail in the chapter 2, methods section.

Chapters 2 and 3 make use of the molecular marker information with a simplified representation (coding), in which only the presence or absence of alleles are coded. Since sugarcane is a polyploid species, molecular marker data could be more informative if dosage information was encoded as well. Nevertheless, this is not a straightforward procedure, and is a question under investigation in sugarcane scientific literature. We approached this topic in chapter 4, taking advantage of the demonstrations already carried out in chapter 2 as a baseline for genomic prediction, and also performed genome wide association tests.

CHAPTER 2

GENOMIC PREDICTION OF CLONE PERFORMANCE IN SUGARCANE

2.1 Introduction

Sugarcane is a crop of commercial importance for many tropical and subtropical regions in the world and ranks 3rd in total value and first in biomass production according to FAO figures (FAOSTAT, 2016). Sugarcane products include sugar (crystallized sucrose, used in the food industry), ethanol (used as biofuel or in beverages), animal feed and electricity (produced from biomass left after juice extraction) (Hoang et al., 2015). Its center of origin lies in the New Guinea island, where domestication of *Sacharum officinarum* took place (Ming et al., 2010). The modern cultivated varieties are derived from crosses between *Sacharum officinarum* and *Sacharum spontaneum*, an undomesticated species that brought significant improvement for disease resistance and for yield into the domesticated genetic background. Both species are polyploids and their interspecific crosses resulted in cultivars with chromosome numbers ranging from 80 to 120, and ploidy ranging from 8 to 14. It has been recognized that its genetic complexity poses significant challenges for determining the molecular basis for many of its traits.

Genetic mapping efforts in sugarcane date back to the 90s (Wu et al., 1992; Grivet et al., 1996; Ripol et al., 1999). Quantitative Trait Loci (QTL) mapping in biparental populations has been reported for traits of agronomic importance such as sugar content (Ming et al., 2001, 2002a; Aitken et al., 2006; Piperidis et al., 2008; Pastina et al., 2012; Liu et al., 2016), biomass and yield components (Hoarau et al., 2002; Ming et al., 2002a; Aitken et al., 2008), flowering (Ming et al., 2002b), and diseases (McIntyre et al., 2005; Raboin et al., 2006; Aljanabi et al., 2007). QTL mapping using association panels has also been reported for many of those traits (Wei et al., 2006; Banerjee et al., 2015; Gouy et al., 2015; Racedo

et al., 2016). With the exception of alleles for brown rust and root rot resistance (McIntyre et al., 2005; Le Cunff et al., 2008), other marker-trait associations have not been confirmed in independent populations.

Identification of QTL presents opportunities for studying the molecular biology of the traits and can be a tool for improvement of traits in breeding programs (Lande and Thompson, 1990). Effect size of QTL and linkage disequilibrium between markers and causal loci are key parameters that determine the usefulness of associated markers. Nevertheless, quantitative traits are inherently difficult to map and often are not confirmed in independent validation experiments. In this context, genomic selection (GS) (Meuwissen et al., 2001) has been proposed as an alternative strategy for improving quantitative traits in breeding programs. In GS there is a shift from statistical estimation of marker causal-loci association and its effect on a trait, to the statistical prediction (genomic prediction, GP) of the genetic value or breeding value for the trait, using all molecular marker information available. Its use in plants has been studied for many crops (reviewed in Heslot et al. (2015)) and was first demonstrated in sugarcane by Gouy et al. (2013).

In the study of Gouy et al. (2013), a population comprised of 334 accessions representing sugarcane diversity from the main growing regions around the world was used to assess genomic prediction for disease resistance, Brix and yield component traits. As pointed out by authors, in spite of the high accuracies obtained, the performance of genomic prediction in the context of a breeding program needs to be validated. It is important to note that sugarcane yield is more effectively measured through the total biomass weight of plots (used to derive the value for tons of cane per hectare (TCH) and Pol, which is a mea-

sure of sucrose content. Those are key traits used in the selection of advanced sugarcane clones in commercial breeding programs (Jackson, 2005; Ming et al., 2010).

Sugarcane is perennial and clonally propagated, with the first harvest taking place one year after planting followed by two or three additional harvests. The process of multiplying clones to obtain enough propagation material for planting a yield plot takes about five years. For evaluation in multiple locations, even more time is required. During those initial years selection is performed on high heritability traits or on less important traits. Also evaluation of more than one harvest is essential as the performance on the first crop (plant crop) does not accurately predict the following harvests (ratoon crops). Taking that into account, the release of a new sugarcane variety will take around 12 years, and identification of new clones as high performance parents takes roughly the same amount of time. In this context sugarcane breeding programs would benefit from accurate predictions of yield related traits, if possible at earlier stages in the breeding program. Consequently, genomic selection could play a role in improving efficiency of a breeding program.

The goal of this project was to determine genomic prediction accuracies using data from the first stage in which yield data were available in a commercial breeding program. Data from two breeding cycles were analyzed, so that the predictions made with data from one year can be compared to the measurements obtained in a different cycle.

2.2 Materials and methods

2.2.1 Plant material and phenotypic data

Plants used in this study came from two breeding cycles performed at CTC - Centro de Tecnologia Canavieira - located in Piracicaba, São Paulo, Brazil. In the *Cycle* '05 a total of 1,732 full sib families were evaluated in a trial with two replicates (from which data of 1,718 families are available) (table 3.1). Each replicate of a family consisted of 68 genetically distinct individual plants, and measurements of Brix (average of sample of plants) and plot weight (Weight, total cane weight in kilograms from a plot) were taken from the plant crop (1st harvest). This step represented the *Stage* 1 of this cycle. For the *Stage* 2 plants, 9,207 clones from *Stage* 1 (belonging to 1,556 different families) were visually selected, clonally multiplied and planted in one replicate, which was evaluated for Brix. *Stage* 3 consisted of 1,233 clones from *Stage* 2 (representing 673 families), which were clonally multiplied a second time and planted with two replicates (table 3.1). For *Stage* 3 the phenotypes available were: Brix, Pol (sucrose content in cane juice as % of biomass, measured by polarimetry), fiber and Weight. For the *Stage* 3 the phenotyping occurred in two periods: for 607 of the clones, measurements were taken early in the season (March - April), and for the other 626 clones, measurements were taken late in the season (July-August), plus three checks were in common between the measurements. Furthermore all traits were measured in the plant crop (1st harvest) and the ratoon crop (2nd harvest).

For *Cycle* '06 a new set of full sib families were used. The evaluation and selection performed in this cycle were similar to *Cycle* '05, with different numbers. In *Stage* 1, data from 1,457 full sib families were available (table 3.1). In

Stage 2, 9,184 clones from 1,266 families were available and in *Stage* 3, 866 clones (420 early and 446 late, plus three checks) corresponding to 490 families were available.

Approximately 25% of the parents used for the crosses in *Cycle* '05 were also used in crosses for *Cycle* '06, but no single biparental cross was repeated in *Cycle* '06. All clones had their pedigree recorded, with an average depth of 10 generations (Atkin et al., 2009). Due to the selection that occurred within a Cycle (from *Stages* 1 to *Stages* 3) the size of families present in *Stage* 3 were variable, as presented in fig. 2.1.

Table 2.1: Field data overview. Number of genotypes, phenotypic records and related statistics are presented, organized by their originating *Cycle*, *Stage* and *Season* if applicable.

Cycle	Stage	Season	Records[a]	Genotypes[b]	in $\boldsymbol{K}$[c]	Families	Females	Males	Total # parents	Planting Year	Harvest Years
'05	1	-	3943	1718	0	1718	605	147	650	2005	2006
'05	2	-	9379	9209	1	1558	580	146	637	2007	2008
'05	3	early	2680	610	534	389	246	69	280	2009	2010/2011
'05	3	late	2768	629	572	443	246	94	293	2009	2010/2011
'06	1	-	3122	1457	0	1457	629	137	681	2006	2007
'06	2	-	9592	9187	2	1269	633	160	699	2008	2009
'06	3	early	1824	423	259	273	198	89	255	2010	2011/2012
'06	3	late	1976	449	397	285	216	88	274	2010	2011/2012
		Total	35284	16805	1755	3322	1240	249	1306		

Note: [a] Number of field records per trait, including replicates and checks. [b] Refers to families when *Stage* 1, clones when *Stage* 2 and 3. [c] Genotyped individuals, includes checks.

2.2.2 Statistical analysis of phenotypic data

Plants were laid out in the field in units called *Sets*. Within each *Set*, *Row* and *Column* information was available describing the relative spatial layout of plants in the field. Each *Set* was divided into super-blocks, each of them being comprised

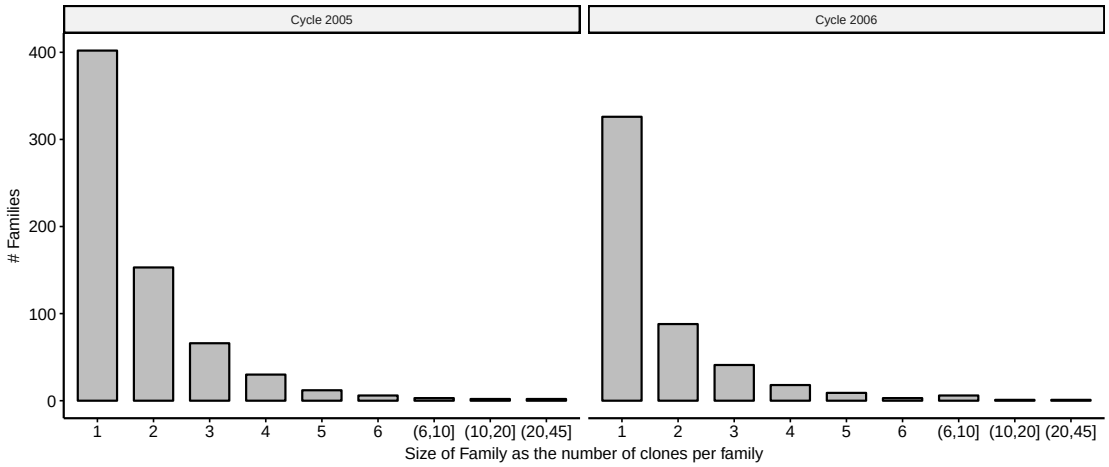

Figure 2.1: Number of clones evaluated per family in *Stage* 3, for both *Cycle* '05 and *Cycle* '06.

of two blocks (in the case of *Stage* 1 and *Stage* 3 datasets) in a Randomized Complete Blocks Design, or one block (in the case of *Stage* 2 dataset). Each individual genotype or family is present in only one super-block, with the blocks being their replicates (two in *Stage* 1 and *Stage* 3, and only one in *Stage* 2). Checks are present across all blocks so that the super-blocks taken together form an Augmented Blocks Design. Taking into account this field configuration, phenotypic measurements were analyzed in a linear mixed model in order to extract adjusted means for each genotype in the study. This was done in two steps. In the first step each *Set* was analyzed independently to determine what spatial correlation structure best fit the data. For each trait and *Set* combination a model was used as defined by:

$$y_{ijk} = \mu + \alpha_i + \beta_{j[i]} + g_k + e_{ijk} \tag{2.1}$$

Where y_{ijk}: trait measurement, μ: overall mean, α_i: effect for *Super-block*

i, $\beta_{j[i]}$: effect for *Block* j nested within *Super-block* i, g_k: effect for genotype k, which is a given clone in the case of *Stages* 2 and 3 datasets, and a cross in *Stage* 1 dataset and e_{ijk}: residual error. Both μ, α_i and $\beta_{j[i]}$ were considered fixed effects, and g_k and e_{ijk} random effects, Normally distributed as $g_k \sim \mathcal{N}\left(0, \sigma_g^2 \boldsymbol{I}\right)$ and $e_{ijk} \sim \mathcal{N}\left(0, \sigma_e^2 \boldsymbol{R}\right)$, with σ_g^2 the genetic variance, $\boldsymbol{I}$ the identity matrix, σ_e^2 the residual variance, and $\boldsymbol{R}$ a correlation matrix for the residual effects. In this this first step, different possibilities for $\boldsymbol{R}$ were tested: the identity matrix so that residuals were assumed independent, first order autocorrelation matrices for *Row* and *Column* (only *Row*, only *Column*, and both), and second order autocorrelation matrices for *Row* and *Column*. All those matrices were parameterized accordingly. Then Akaike's Information Content (AIC) was used to select the best model for that *Set* and trait combination.

In the second step, for each trait a model with the following shape was fit:

$$y_{ijkl} = \mu + \delta_i + \alpha_{j[i]} + \beta_{k[i,j]} + g_l + e_{ijkl} \tag{2.2}$$

Where terms already present in eq. 2.1 had the same meaning here, and the term δ_i was the fixed effect for each *Set* i. This model was fitted to all data, across *Set*s. The matrix $\boldsymbol{R}$ had a different formulation too, being equal to $\boldsymbol{R} = \bigoplus_{i=1}^{n} \boldsymbol{R_i}$, the direct sum of the best matrix $\boldsymbol{R_i}$ for *Set* i in the first step. The parameters that defined $\boldsymbol{R_i}$ were fitted again in this last model.

All model fitting at this point was run in statistical package R (R Core Team, 2016), using Restricted Maximum Likelihood (REML) in `Asreml` package (Gilmour et al., 1995). The Best Linear Unbiased Estimates (BLUPs) obtained from models fitted using eq. 2.2 were subsequently used as the total genetic

value of each genotype for a given trait.

Clone based broad sense heritability estimates were obtained for each *Set* as:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2/r} \tag{2.3}$$

Where σ_g^2 and σ_e^2 were the variance estimates obtained in the first step described above and r is the number of replicates in each experiment, which was equal to two for *Stage* 1 and *Stage* 3 datasets, and equal to one for the *Stage* 2 dataset.

2.2.3 Genotypic data

Molecular marker data were obtained only for clones in *Stage* 3 from both *Cycle* '05 and *Cycle* '06. Genotyping was carried out by Rapid Genomics, Gainesville, FL, USA, using methodology similar to Bundock et al. (2012) and Song et al. (2016). In brief, DNA was extracted from leaf samples and fragmented through sonication, generating random length fragments from the genome. A set of 40,000 probes of length 120 base pairs were designed from publicly available sugarcane expressed sequence tags (Vettore et al., 2003), and then used for capturing the DNA fragments that hybridized to the probes, yielding genome complexity reduction to fragments that show sequence similarity to the probes. Captured fragments were sequenced and their sequencing reads were aligned back to the probe sequences. Assuming the set of reads that align to a given probe belong to the same locus in the genome, but could originate from different alleles at the given locus, the set was used for SNP calling, taking into account the se-

quencing quality and variation in single bases among those aligned sequences. This procedure resulted in an initial set of 245,923 putative SNPs from 1,882 different clones or cultivars. These SNPs came from 31,808 different probes so not all probes resulted in detected polymorphic sites, with an average of 7.73 SNPs per probe. Marker genotypes were coded 0, 1 and 2 for observation of only the reference SNP allele from the probe sequence, both alleles and only the alternative allele, respectively in the sequencing reads.

Because of the polyploid and heterozygous nature of sugarcane a low dosage allele may be missed in the genotyping if it is derived from a small number of reads. For this reason, criteria for filtering the initial set of SNPs were developed. Using the repeatability of genotype calls for 19 replicated DNA samples, it was observed that 50 reads per individual genotype minimized the difference between replicates but resulted in a drastic reduction in the number of markers, compromising genome coverage. The final filtering criteria included a minimum of 30 reads per individual genotype, less than 75% missing data, more than 5% of the clones having an alternative allele, and less than 25% missing clone data. Using these criteria, the final set of markers had 54,675 SNPs in 1,778 different clones and cultivars. And those SNPs came from 12,125 different probes, with average of 4.51 SNPs per probe. The filtering process generated missing data, which was imputed using the weighted k-nearest neighbors imputation (kNNI) method with $k = 4$ (Troyanskaya et al., 2001; Rutkoski et al., 2013), and using the correlation between molecular markers, instead of Euclidean distance to determine the k nearest markers.

Finally SNP data were used to obtain a genomic relationship matrix ($\boldsymbol{K}$) as

in:

$$\boldsymbol{K} = \frac{\boldsymbol{M}\boldsymbol{M}^t}{k} \tag{2.4}$$

Where $\boldsymbol{M}$ is the mean centered molecular marker matrix with SNP information in columns and individuals in rows. The scaling constant k does not impact prediction accuracies, but here it was chosen to be $mean(diagonal(\boldsymbol{M}\boldsymbol{M}^t))$ so that the $\boldsymbol{M}$ diagonal has comparable scale to the pedigree derived relationship matrix $\boldsymbol{A}$ which also has a mean diagonal value close to 1.

The molecular marker data were further explored in two ways. First, in order to observe the overall genome coverage and distribution of molecular markers, the probe sequences were aligned to the sorghum genomic sequence (Paterson et al., 2009) and statistics about the counts of markers per megabase segment were registered. Also, in order to observe the relation between linkage disequilibrium and marker coverage, the association between pairs of markers belonging to probes that aligned to the same chromosome were computed for one random clone per full sib family, using the Pearson correlation as the statistic. Second, the population structure due to overall molecular similarity among genotyped sugarcane clones was analyzed through Principal Component Analysis, obtained by the single value decomposition of the $\boldsymbol{K}$ matrix.

2.2.4 Genomic prediction and validation

Genomic predictions were performed using a genomic BLUP method (Meuwissen et al., 2001) with two kernels, one for pedigree data and another for molec-

ular marker data, as represented by the linear mixed model:

$$y_i = \mu + u_i + v_i + e_i \tag{2.5}$$

Where y_i uses the BLUP estimates from the step in eq. 2.2 for a given trait, μ is overall mean, and the remaining terms are random effects with distributions as $u_i \sim \mathcal{N}(0, \sigma_u^2 \boldsymbol{A})$, $v_i \sim \mathcal{N}(0, \sigma_v^2 \boldsymbol{K})$, $e_i \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I})$, with $\boldsymbol{I}$, $\boldsymbol{A}$ and $\boldsymbol{K}$ as mentioned in the previous sections.

In order to compare predictions based on pedigree relationships with predictions based on molecular marker relationships, alternative configurations of this model using either the u_i term or the v_i term were also tested.

Assessment of quality of predictions was performed using cross-validation within *Cycle* '05 dataset and through prediction across cycles. In the former setting, records were randomly assigned to 10 subsets, and each subset was used for validation (so their phenotypic value y_i was assigned as missing) of the model trained based on the remaining sets. This process was repeated six times with different random subsetting assignments (i.e. a 10 fold cross validation, with 6 replicates). In the latter setting, all the data from *Cycle* '05 were used as a training set, and *Cycle* '06 as validation set.

In each step prediction accuracy was estimated as the average of the Pearson correlation between the predicted values (u_i + v_i) and the observed values (y_i). This process was repeated for each trait considered in this study.

2.3 Results

2.3.1 Molecular information from sugarcane clones

A sugarcane genome reference sequence is not yet available, but coverage can be assessed using the sorghum genome as reference, which has more than 90% similarity to the sugarcane genome (Ming et al., 1998; Wang et al., 2010). The probe sequences, which were the origin of the detected SNPs, were aligned to the sorghum genome and the number of SNPs present in each genomic position was plotted as histograms in fig. 2.2. Despite the reduction in the number of markers, the genome was well covered. This analysis has inherent limitations because of the lack of a sugarcane genome reference sequence and also due to the polyploid/aneuploid nature of sugarcane in contrast to diploid sorghum. Also probes might be tagging multiple alleles at the same locus in homologous chromosomes and in different loci in the genome (paralogous regions). Nevertheless, using the alignment of probe sequences to the sorghum genome, we assessed the distance and association between neighboring markers in the final set of filtered SNPs. On average there were 4.51 SNPs per probe, and the median distance between neighboring probes is 10,717 base pairs (the distribution of distances was skewed and the average was 54,366 base pairs). The average absolute correlation between markers in neighboring probes was 0.08. Given that the average absolute correlation between markers within the same probe was 0.12, there was a relatively high association between markers across neighboring probes.

The Principal Component Analysis (PCA), obtained from the single value decomposition of the covariance matrix between genotypes (due to similarity in

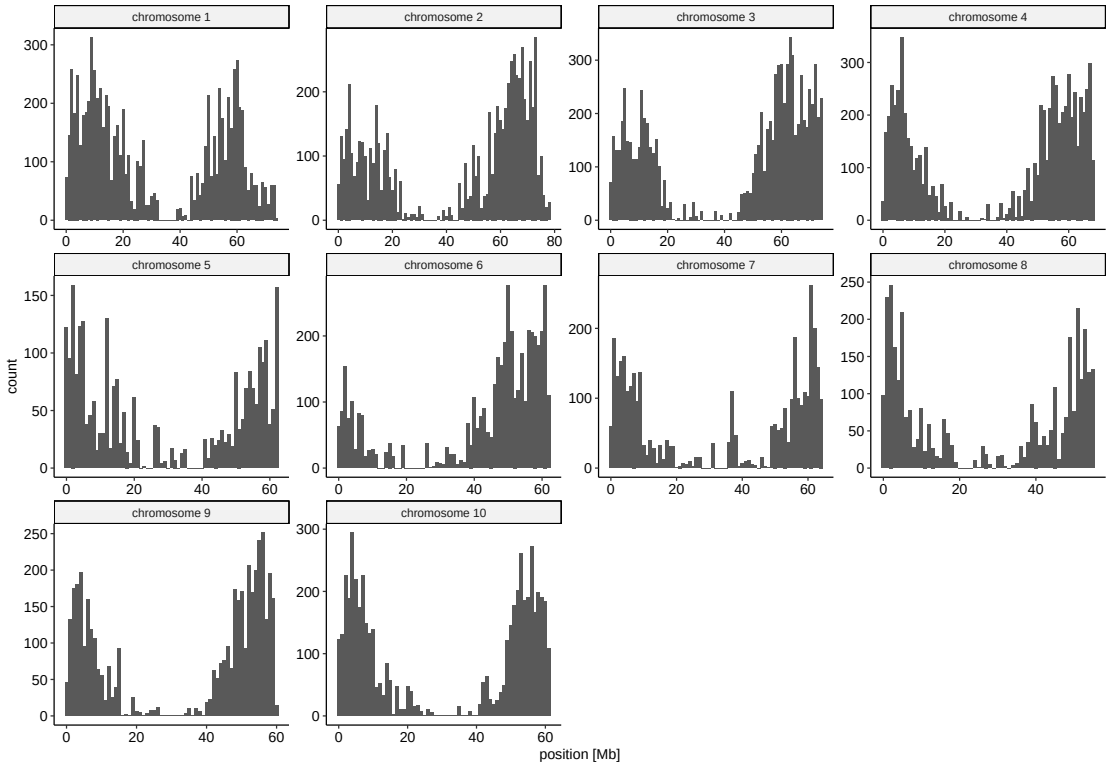

Figure 2.2: Alignment of markers to the sorghum genome (Paterson et al., 2009) shows overall genomic coverage of markers. Histograms of the count of probes in bins of 1Mb is shown, for each chromosome.

the SNP data), was used in the plot of fig. 2.3, with panels **A**, **B** and **C** showing the same plot but with different color schemas to distinguish which cycles the clones belong to (panel **A**), or female parent (panel **B**) and male parent (panel **C**) of the clone. The clones that originated from *Cycles* '05 and '06 show clear overlap (fig. 2.3, panel **A**). The first and second (largest variance) components from the PCA explained 17% and 10%, respectively, of the total variance. The groups identified the clones that were progeny of the most frequent male parents in the selected population that make up *Stage* 3 (fig. 2.3, panel **C**). Therefore, the population stratification observed in the PCA plots are more related to close family relationships than to differences among the breeding cycles. One

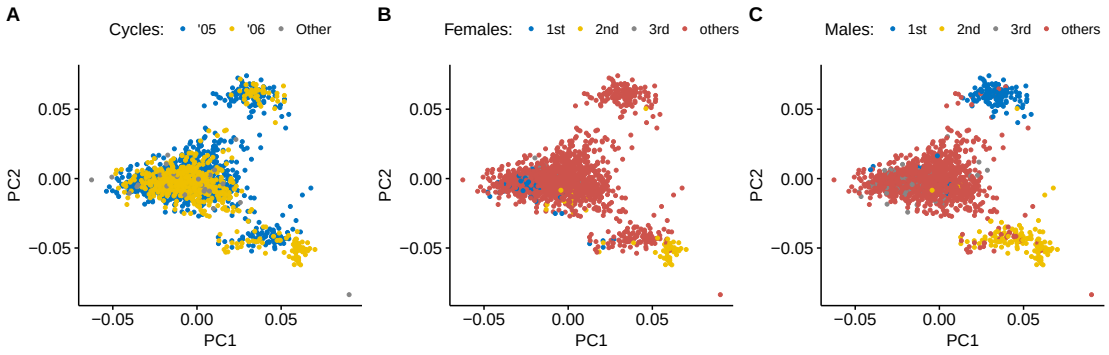

Figure 2.3: PCA plot from the first and second components obtained by eigen decomposition of the molecular marker matrix, explaining 17% and 10%, respectively, of the total variance. Each dot indicates one genotyped clone. The three panels present the same components, with dots colored to show different information from the clones: **A**) colors indicate the cycle that the clone belongs to, with the "Other" category applying to 26 checks and parents also genotyped; **B**) colors indicate the female parent of the clones (the 3 most common female parents had 59, 53 and 40 clones); **C**) colors indicate the male parent of the clones (the 3 most common male parents had 149, 131 and 97 clones).

should note that this is a breeding population, and sugarcane breeding reuses successful parents through many cycles and that may explain the lack of population structure and overlap between clones from different cycles. The PCA plots demonstrate that the SNP data still retain biologically relevant relationships between the clones.

Pairwise relationship estimates were derived from molecular markers and pedigree data (fig. 2.4). Despite the deep pedigrees, information can be sparse in some cases due to the common use of polycrosses where only information regarding the female parent is recorded. In general there is modest agreement between the markers and pedigree relationships with a correlation of 0.63 (correlating off-diagonal elements from matrices $\boldsymbol{A}$ and $\boldsymbol{K}$). For marker derived

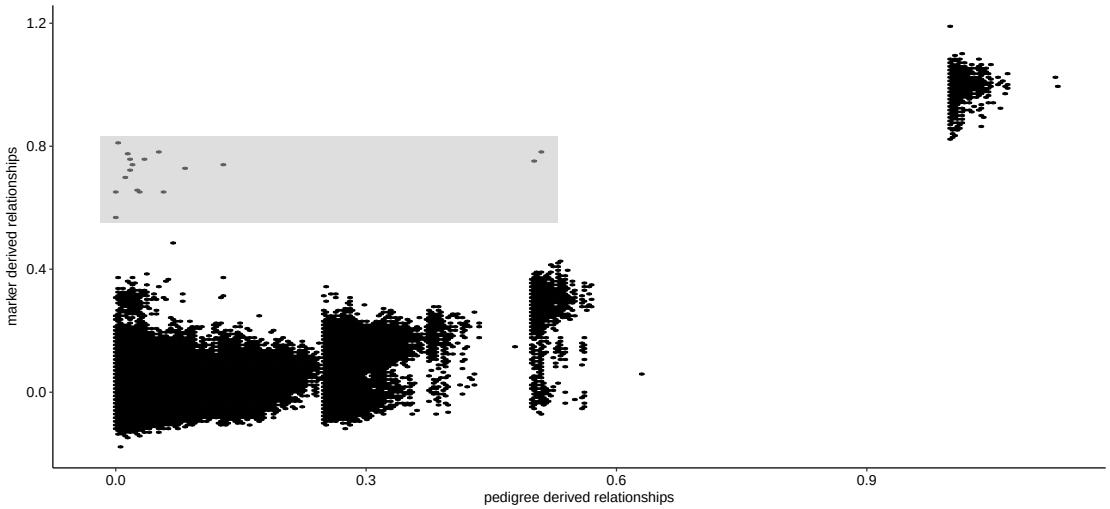

Figure 2.4: Relationship coefficients derived from pedigree data plotted against the scaled relationships derived from molecular marker data. Correlation between vertical and horizontal axes is 0.63, when excluding the diagonal elements in the relationship matrices (dots above 0.9 in the horizontal axis). Shaded region indicates pairs of genotyped clones considered outliers in their marker relationships and then excluded from subsequent analyses.

relationships below 0.2, the corresponding pedigree derived relationships show a greater spread, whereas the opposite can be observed for pedigree relationships close to 0.5 and 1 where marker relationship had greater spread. It was observed that 17 pairs of genotyped clones had marker relationships values at the range of observed values for relationships of duplicated samples, they were also outliers in the distribution of marker relationships between pairs of different clones. The corresponding 34 clones were removed from subsequent analyses, due to the possibility of mislabeling in the genotyping or mis-assignment of plants to plots in the field.

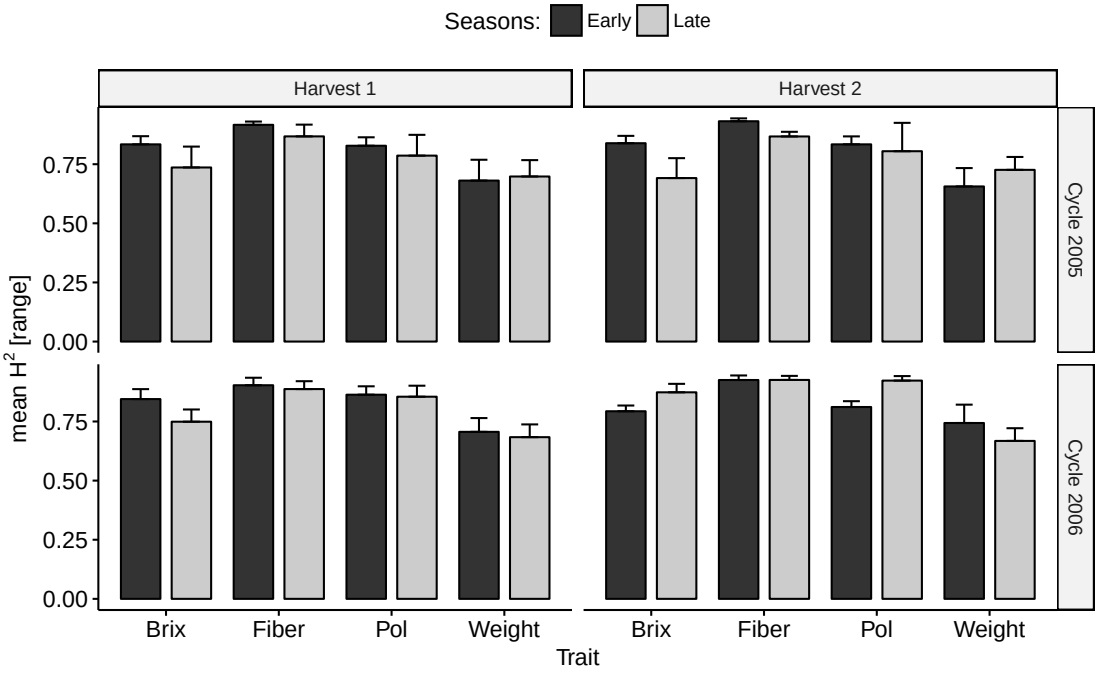

Figure 2.5: Broad sense heritabilities (H^2) for the different traits evaluated for clones at *Stage* 3 from the breeding program. The mean value across different *Set*s in each combination of harvest, cycle, season and trait is shown. Each bar is an average of 5 to 14 different sets, with the range of values depicted in the error bars.

2.3.2 Prediction accuracies

The clone based heritabilities (H^2, Broad Sense Heritability) for the four different traits were measured at *Stage* 3, both for *Cycle* '05 and *Cycle* '06 (fig. 2.5). The original estimates, using eq. 2.3, were obtained for each *Set* in the field, so we observe variation for the estimates, even on a given *Harvest* and *Cycle*. There were non-significant differences in the estimates across *Cycle* and across *Harvest* and *Season*. Heritability of Weight had the lowest average heritability (0.69), the sugar related traits slightly higher values, Brix (0.78) and Pol (0.83), and Fiber the highest (0.90).

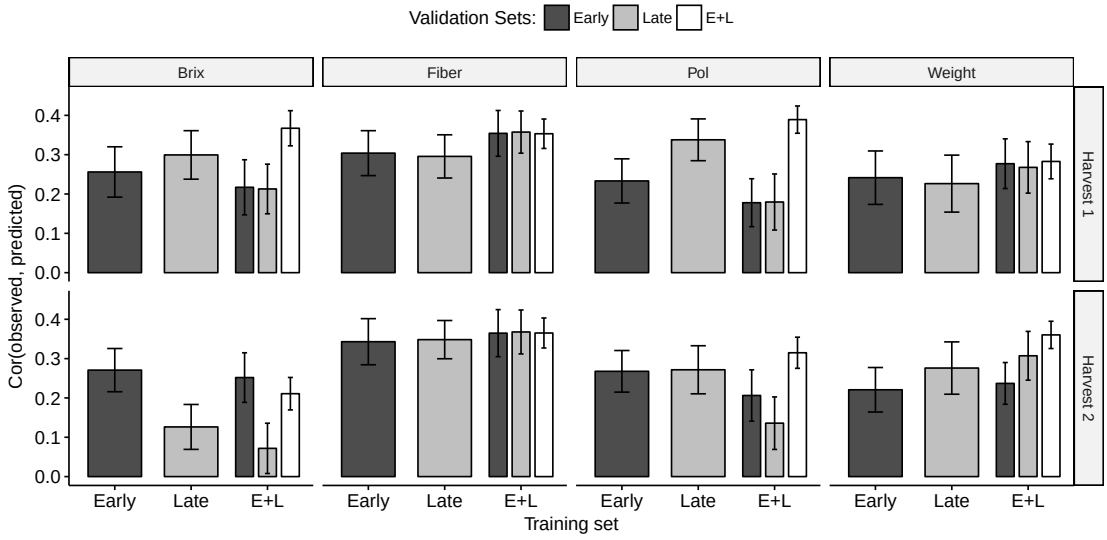

Figure 2.6: Prediction accuracy for genomic selection for clones at *Stage* 3 from *Cycle* '05. Values shown are mean values of 6 replications of 10 fold cross validation for each trait, with error bars being the respective standard deviation. Clones were evaluated at *Early* or *Late* seasons, so the training and validation sets can be comprised of either of those sets or both seasons combined (E+L). Model for prediction used 2 kernels, with matrices $\boldsymbol{A}$ and $\boldsymbol{K}$.

The prediction accuracies for clones at *Stage* 3 were calculated as an average of a 10 fold cross validation with 6 replicates, with values in the range of 0.07 to 0.39 (fig. 2.6). The datasets were evaluated for *Early* and *Late* seasons, so that prediction could be done for the same season keeping data from the same given season in the training set (the two leftmost bars in each cell of fig. 2.6), or one can combine *Early* and *Late* data in the training set and use it to predict only *Early*, only *Late* or both *Early* and *Late* cases (three rightmost bars in cells of fig. 2.6). For a given trait, *Early* and *Late* and both harvests 1 and 2 were similar. Consistent with heritabilities, prediction accuracies were higher for Fiber and lower for Weight.

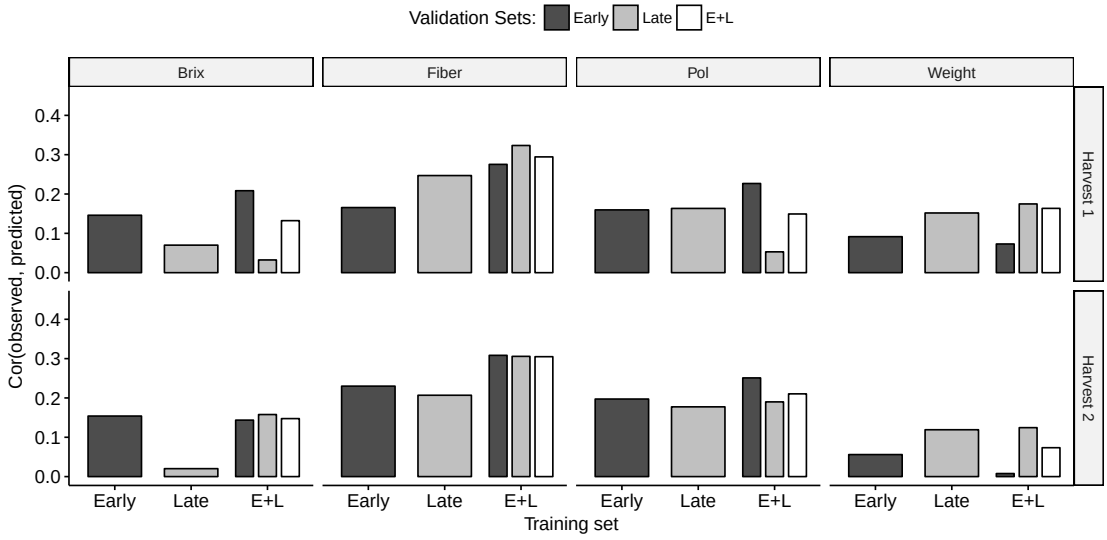

Figure 2.7: Prediction accuracy for genomic selection for clones at *Stage* 3, using *Cycle* '05 as training set and *Cycle* '06 as validation set. All traits are shown for both harvest 1 and 2. Clones were evaluated as *Early* or *Late* seasons, so the training and validation sets can be further subdivided into either of those sets or both seasons combined (E+L). Model for prediction used 2 kernels, with matrices $\boldsymbol{A}$ and $\boldsymbol{K}$.

Model training based on *Cycle* '05 was used to predict clone values in *Cycle* '06 resulting in prediction accuracies ranging from 0.01 to 0.32 (fig. 2.7). As in fig. 2.6, accuracies across *Harvest* and *Season* were similar, but for Fiber there was a consistent increase in accuracy when using *Early* and *Late* data together (leftmost bars in fig. 2.7) probably because of increased training population size. Overall Fiber had higher accuracies than the other traits. For *Early Season*, Weight had lower accuracies, and for *Late Season* Brix values were lower. There was a decrease in accuracies in comparison to the cross-validation results (shown in fig. 2.6).

2.3.3 Molecular marker vs. pedigree prediction

All results presented in figs. 2.6 and 2.7 were obtained using kernels for both matrix $\boldsymbol{A}$ (term u_i) and $\boldsymbol{K}$ (term v_i) in eq. 2.5. Models were fit using only one of those terms and results are presented in figs. 2.8 and 2.9, including results for all traits in the same combinations of harvests and seasons as in previous results. For cross-validation using only *Cycle* '05 data, most of the predictions using only molecular marker data (vertical axis in fig. 2.8, panel **A**) were higher than predictions only using pedigree data (horizontal axis in fig. 2.8, panel **A**). Panel **B** in fig. 2.8 has similar results but using pedigree and molecular marker data together (on the vertical axis). Changes in accuracies from panel **A** to **B** indicate that improvement in prediction accuracies can be obtained by inclusion of a kernel for pedigree information, even as in most cases the kernel for molecular markers (v_i, using $\boldsymbol{K}$) had higher weight for predictions. In fact 97% of all cross-validation iterations had a higher value for the variance estimate for the term v_i than for the term u_i in eq. 2.5, even with 67% of them having non-zero variance estimates for the term u_i (related to matrix $\boldsymbol{A}$).

For the case where prediction is done using *Cycle* '05 data and validation using *Cycle* '06 data (fig. 2.9), a larger proportion of the results showed higher accuracy for prediction using only pedigree data than in the cross-validation case. For both the comparison of prediction using only molecular markers versus only pedigree (panel **A**, fig. 2.9) and the comparison of prediction with both molecular markers and pedigree versus only pedigree (panel **B**, fig. 2.9), 32% of the trait-season-harvest combinations had accuracies higher for predicting using only pedigree. For the predictions using both kernels, 67% of cases had a non-zero variance estimate for term u_i (related to pedigree), whereas variance

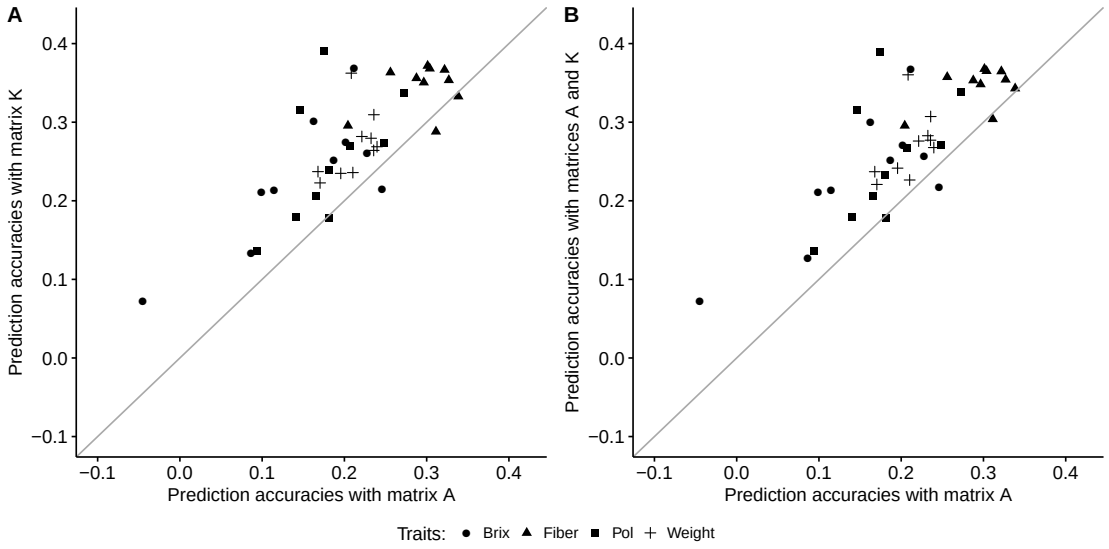

Figure 2.8: Comparison between prediction using only molecular markers (with matrix $\boldsymbol{K}$) and pedigree (with matrix $\boldsymbol{A}$), for cross validation using *Cycle* '05 data. For a given trait, all combinations of *Early* and *Late* seasons that are present in fig. 2.6 are shown here together. Pedigree prediction accuracies are plotted on the horizontal axis (in both panels **A** and **B**), and the vertical axis has prediction accuracies only with markers (panel **A**) and markers and pedigree (panel **B**). The diagonal line indicates value combinations where accuracies would be the same.

estimates for the term v_i were all non-zero.

2.4 Discussion

In this study, genomic prediction for an applied sugarcane breeding program was evaluated for predicting total genetic values of clones. Due to costs associated to clonal multiplication and phenotypic measurements in large scale trials, a sugarcane breeding program accumulates costs and time along several stages of evaluations. Procedures and technologies that can allow accurate selection

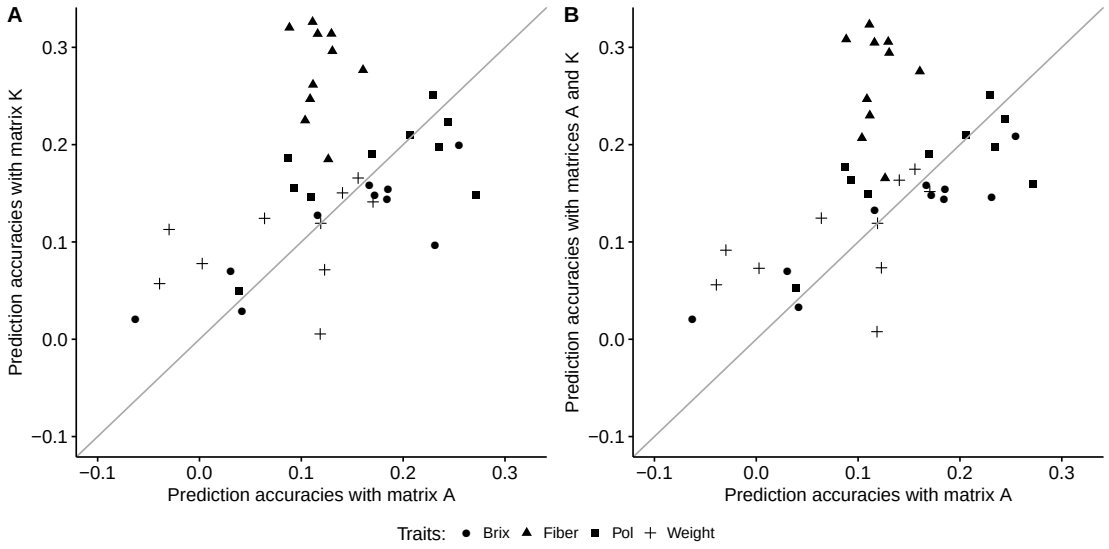

Figure 2.9: Comparison between prediction using only molecular markers (with matrix $\boldsymbol{K}$) and pedigree (with matrix $\boldsymbol{A}$), using *Cycle* '05 data for training and *Cycle* '06 for validation. For a given trait, all combinations of *Early* and *Late* seasons that are present in fig. 2.7 are shown here together. Pedigree prediction accuracies are plotted on the horizontal axis (in both **A** and **B**), and the vertical axis has prediction accuracies only with markers (panel **A**) and markers and pedigree (panel **B**). The diagonal line indicates value combinations where accuracies would be the same.

at earlier stages are promising and the costs associated with those technologies might be offset by the reduction in phenotyping costs and breeding cycle time, resulting in increased genetic gain.

2.4.1 Genome wide molecular marker data

Recent studies have demonstrated the use of Genotype-by-Sequencing (GBS), and variants, as a promising technology for genome wide molecular markers for sugarcane (Bundock et al., 2012; Song et al., 2016; Yang et al., 2017). This

study used GBS technology with targeted enrichment of genomic regions to genotype all clones in a sugarcane breeding program.

Determination of the distribution of the markers across the sugarcane genome is difficult because the full genome sequence of sugarcane is not yet available. Consequently, we used the genome sequence of the closest relative which is sorghum. Probe sequences that were used to develop the molecular markers were well distributed throughout the sorghum genome, covering all arms in all chromosomes. Because there is more than 90% similarity between sorghum and sugarcane genomes (Ming et al., 1998; Wang et al., 2010), it was inferred that markers came from all sugarcane chromosomes. As sugarcane is a polyploid with up to 10 homologous chromosomes, coverage shown in fig. 2.2 could be sparser than predicted. On the other hand, it is also possible that not all homologous chromosomes have different alleles at a given locus resulting in multiple dosage alleles, and this might limit the sparsity. With SNP data in an autopolyploid like sugarcane, it is possible that a SNP allele tags different haplotypes at a locus. These considerations likely contribute to the low, mean absolute correlation between different markers within the same probe (0.12). Because this statistic is similar to the average correlation between markers from nearby probes (0.08), it suggests that the coverage of the genome was adequate for this kind of molecular marker system. It has been suggested through simulations that an R^2 of 0.15 (or corresponding correlation of 0.39) would be necessary for a high heritability trait (Lorenz et al., 2011; Calus and Veerkamp, 2007). One path to achieve a higher correlation between markers would be a marker system that is able to exclusively tag the different alleles possible for a locus in sugarcane. This would improve the molecular marker effectiveness in tagging the existing allelic diversity.

2.4.2 Prediction of clonal performance

The availability of dense marker information enables clonal prediction. Applying the GBLUP modeling, we performed predictions focusing on clonal total value prediction only taking into account additive genetic effects to predict their total genetic values. Those values were compared to the total genetic values of *Stage* 3 clones when estimating the prediction accuracy.

Using cross validation, we obtained prediction accuracies ranging from 0.07 to 0.39 (fig. 2.6), with the highest values for the highest heritability traits. In Gouy et al. (2013), the mean prediction accuracy for cross validation within the same panel ranged from 0.11 to 0.62, using 10 traits measured in two panels of 167 sugarcane accessions (totaling 334 accessions). The only trait in common with the current study was Brix, which had a median prediction accuracy of 0.47 and 0.62 for the two panels (the remaining traits had median accuracies between 0.11 and 0.50), and broad sense heritability between 0.83 and 0.88. Despite the heritabilities for Brix reported here being similar (average 0.78, range from 0.53 to 0.90), the prediction accuracy was lower with an average of 0.26. Stalk Number (SN) was another trait used in Gouy et al. (2013), which was not measured here but is highly correlated with Weight (Brown et al., 1969; Aitken et al., 2008). Comparing SN from Gouy et al. (2013) and our Weight data, SN showed prediction accuracy with medians of 0.25 and 0.46 in their two panels, and heritabilities of 0.8 and 0.9. Whereas in the present study, Weight had average of 0.27 for prediction accuracy, with mean heritability 0.69 (ranging from 0.47 to 0.86). Therefore SN in Gouy et al. (2013) was predicted with higher accuracy than Weight here, which can possibly be attributed to the higher heritability of SN in comparison to Weight.

One important difference between the current project and Gouy et al. (2013) was the panel composition, which in their case was a sample of cultivars and breeding materials from over 30 different breeding centers (Costet et al., 2012), and so it was expected to be more diverse and less adapted to the evaluated environments than the breeding material used in the current study. In the PCA analysis in Gouy et al. (2013) the first 3 principal components explained about 10% of variance in the relationship matrix derived from molecular markers, whereas the current study obtained 34% for the same statistic, which is evidence for a population with more structure (Patterson et al., 2006). Greater structure could also be associated with lower prediction performance because markers in training versus validation sets may be found in different linkage phase to QTL in divergent plant material (Riedelsheimer et al., 2013). The biparental populations used in Riedelsheimer et al. (2013) showed complete separation in PCAs, which was not observed in Gouy et al. (2013) and was partially present in the current study. We did not observe increased prediction accuracies when restricting training and validation sets to only the portion of clones in the larger cluster in the PCA plot (excluding clones that are progeny from the 2 most common male parents, fig. 2.3 panel **C**). This may indicate that diversity was not high enough to cause different linkage phase between markers and QTL in the diverse material used in Gouy et al. (2013), but instead sampled different combinations of independent QTL to allow better model training. This would be consistent with the high polyploid nature of sugarcane and its breeding history (Jackson, 2005).

2.4.3 Effect of GxE in prediction accuracies

The clonal prediction using cross validation assessed the prediction accuracy taking into account data from a single year of evaluation and did not account for Genotype by Year (GxY) variation. Even if we combined the two available harvests from subsequent years, this GxY variation would not be taken into account properly because the two harvests (plant crop harvest and ratoon crop harvest) were from the same planting. The harvests were from two different plant developmental stages so not only GxY variation but also variation due to developmental differences were impacting the differences between measurements in those two years. This should also be taken into account when interpreting the results from the heritability for clones (fig. 2.5), as this heritability reflects the accuracy in phenotypic measurements in a given year and harvest without sampling possible variation due to GxY. For this reason, one would expect the heritability estimates to be more comparable to the prediction accuracies in cross validation (fig. 2.6) than the prediction accuracies for clones across cycles (fig. 2.7). Nevertheless, the clonal predictions (fig. 2.6) were lower than the heritability estimates, and prediction accuracies across cycles (fig. 2.7) were even lower, emphasizing these GxY considerations.

The predictions across cycles simulates one possible scenario for application of GS in a breeding program, where phenotypic records on clones from a previous cycle were used to train a model that is then used with genotypes of a new set of clones for selection. It also shows the possible results if all, or part, of this new set of clones under selection were phenotyped in parallel and then used in comparison to the GS. Clearly it would be not properly evaluated since variations due to GxY would have not been taken into account in the training

set, as was the case here. Also the validation set, in this case, the population under selection, or the subsequent cycle, being only evaluated in a single year, will bias the evaluations compared to measurements taken from multiple years. Genotype by Environment (GxE) interaction is a classical theme in plant breeding research and the GS literature has been exploring its implications. Previous results showed that prediction models can perform poorly if training sets and validation sets come from different environments (Resende et al., 2012; Heslot et al., 2014; Jarquín et al., 2014), as was the case in our study. This shows that a prediction and validation scheme would benefit from multiple years of field trials for both training and validation sets. This was partially demonstrated in the across-panel predictions from Gouy et al. (2013), where both panels were phenotyped for several years and the resulting prediction across panels (which is also a prediction for different locations and years) resulted in a small reduction of accuracy in comparison to the cross validation.

Predictions for *Stage* 3 clones using both molecular markers and pedigree allowed the comparison of their prediction accuracies. Using data from the same environment (fig. 2.8) there was a consistent improvement in prediction using molecular markers, but not for prediction across cycles (fig. 2.9). These results suggest that the only advantage to using molecular markers was for *Stage* 3 clone predictions. In cases where pedigree records are not available or for prediction within full sibs, the molecular markers have a clear advantage.

2.5 Conclusions

Considerable levels of accuracy for GP have been observed for sugarcane before (Gouy et al., 2013), but a study based on a breeding population, in a commercial setting that observes the impact of the limitations and advantages brought by the breeding program was lacking and was addressed in this study. The prediction accuracies were in general lower than observed before, and taking into consideration that different traits were used in the current study, this difference in accuracy can be used in planning future deployment of GS for sugarcane breeding. Genotype by Environment interactions can have an important impact in prediction accuracies, as reported here, and should also be taken into account. Fiber, which was a trait measured with the highest H^2 in our field trials, provided the highest genomic prediction accuracies. It would be interesting to check whether field trials designed to provide higher H^2 to the other traits, such as Weight, would obtain higher prediction accuracies in GP as well.

Given the small improvement in accuracy in comparison to prediction using pedigree, breeding schemes that focus on within family selection would obtain more benefit from the use of molecular marker based predictions.

CHAPTER 3

GENOMIC PREDICTION OF PARENTAL BREEDING VALUES IN SUGARCANE

3.1 Introduction

Adoption of landraces was the norm in sugarcane breeding for centuries, before breeding of new varieties started in Java and Barbados in 1888 (Ming et al., 2010). Crosses between the cultivated, sugar-rich species, mainly *Saccharum officinarum* L., and the wild relative, *Saccharum spontaneum* L., that provided resistance to diseases and yield gains, proved successful in the early breeding efforts. Subsequently the focus shifted to breeding and selection among the hybrids themselves (Ming et al., 2010). Sugarcane exhibits severe inbreeding depression preventing the development of inbred lines but is easily clonally propagated, so sugarcane cultivars are clones selected among progenies of heterozygous parents.

Planting true seeds from crosses is followed by selection among seedlings in a series of steps with harvests taking place between 8 and 12 months after planting and then individual plants are clonally replicated. Broad sense heritability at the individual plant level is low, estimated at 0.10 or 0.17 for cane yield in different locations (Skinner et al., 1987) due to large environment and competition effects. Selection at the first stage normally happens at the family level (Bressiani et al., 2005; Kimbeng and Cox, 2003; Stringer et al., 2011; Zhou and Lichakane, 2012), which shows higher broad sense heritability of up to 0.75 for cane yield (Skinner et al., 1987). It is followed by clonal multiplication of the selected seedlings from the best families in subsequent stages, in order to allow evaluation of clones in large plots and in multiple locations. Sugarcane is also perennial, so that after a first year plant harvest, plants are regrown (ratoon crop) allowing several harvests with decreasing harvest yield. The ratio between yield on the first harvest and subsequent ones is genotype dependent

(Milligan et al., 1996; Zhou and Shoko, 2012), so that the proper assessment of genotypes requires multiple year evaluations. This factor combined with a low rate of multiplication for clonal propagation, and the logistics involved in evaluating a large number of seedlings results in a selection cycle of 12 to 15 years (Skinner et al., 1987; Kimbeng and Cox, 2003) before new parents can be selected for the next cycle.

Through the selection cycles, gain from selection is directly proportional to the accuracy in which parents are selected, expressed as the narrow sense heritability, and inversely proportional to the time between cycles (breeder's equation). For this reason, the selection of parents to cross is an important task, that impacts the progress of the breeding program.

New parent candidates are commonly selected among the best clones at the later stages of the clonal evaluation process. Many of the important sugarcane traits, like Brix (related to sugar content), fiber content, and rust resistance are known to be quantitatively inherited and controlled by additive genetic effects (Hogarth, 1987). Nevertheless cane yield and its components are known to have important non-additive inheritance (Hogarth, 1987). Selection of parents based on their phenotype is then unreliable, requiring the evaluation of progenies of the parents (Hogarth, 1987; Stringer et al., 2011). One approach to reducing the cycle time would be to evaluate candidate parents in polycrosses or in a few bi-parental crosses, and select the best ones based on the phenotype of the progenies in the first stage (family evaluation), or subsequent stages (clonal evaluation). The selected clones would be reused in subsequent years in a larger number of crosses. More recently, Best Linear Unbiased Prediction (BLUP) has been used for the Breeding Value (BV) estimation of sugarcane clones (Stringer

et al., 1996; Kimbeng and Cox, 2003; Atkin et al., 2009; Stringer et al., 2011). With the BLUP method, data from relatives are taken into account as well as historical data already available from previous cycles can be incorporated to improve BV estimations (Atkin et al., 2009; Dawson et al., 2013).

In order to make use of data from relatives in BV estimation through BLUP, pedigree or marker data can be employed. Pedigree data can be incomplete or inaccurate because of recording mistakes or unintended self-pollination. Estimation of coefficients of coancestry from pedigrees is based on assumptions that do not hold in breeding programs, such as all genotypes originating from the same unrelated base population, which is assumed to be under Hardy-Weinberg equilibrium. The absence of complete pedigree records from the genotypes under consideration, to this base population would introduce biases to the variance estimates using BLUP (Piepho et al., 2008). This can be impactful in sugarcane even when complete records are available, due to the common use of several males in polycrosses in sugarcane breeding programs so that the male parent is unknown. The use of molecular marker data has been proposed to complement or substitute for the pedigree records in plant breeding (Munoz et al., 2014). The use of molecular markers also allows the differentiation of siblings. With the use of molecular marker data the segregation of alleles in crosses can be tracked, and used for estimation of BVs.

The use of pedigree and molecular information can be leveraged together in genome-wide association studies where pedigree information is used for population structure control (Yu et al., 2006), or in determining the relationship between individuals (Crossa et al., 2010; Juliana et al., 2017). In this latter case, both the pedigree information and the molecular information are used to com-

pute relationship matrices (Habier et al., 2007; Endelman and Jannink, 2012), that can be used together when fitting BLUP models in a multiple-kernel setting, or in a single kernel approach, where both matrices are combined (Legarra et al., 2009). This latter approach, which involves the so called $\boldsymbol{H}$ matrix, brings the possibility of computing relationship matrices that expand the molecular computed relationships to individuals that were not themselves genotyped (Legarra et al., 2009).

The use of molecular marker data in sugarcane has been reported for Quantitative Trait Loci (QTL) mapping, reviewed by Zhang et al. (2013). More recently its use in Genomic Selection (GS) has been reported (Gouy et al., 2013; Brum et al., 2018). In these previous reports the prediction of traits was evaluated in the context of clonal prediction, with the aim of finding the best performing clones on the basis of their total genetic value. Genomic Selection (Meuwissen et al., 2001) has been studied in this context as well as in the selection of parent candidates (Gaynor et al., 2017). In the context of prediction of varietal performance (or the equivalent of clonal performance in the sugarcane context), the total genetic value of a genotype is the goal of prediction, but for predicting parental performance, the focus shifts to the additive value only, as it is the component of the genetic value that is inherited.

Due to the structure of a sugarcane breeding program, historical phenotypic data is comprised of measurements taken on family level at the first stage and on the clonal level in more advanced stages. In previous studies on BV estimation for sugarcane, only the data at the family level has been used. The use of clonal performance of progenies from a parent has been used as an improvement on the BV estimation, but without the use of BLUP estimation (Hogarth, 1987; Skin-

ner et al., 1987). Due to its ability for analyzing unbalanced data (Piepho et al., 2008), the BLUP method would be suitable for incorporating both family data as well as clonal data in BV estimation.

The goal of this study was to evaluate GS for parental prediction for sugarcane, using data from a commercial breeding program. Data from both families and clones selected from those families were available, and the incorporation of both data was evaluated. Genomic Selection in this context was performed using the Genomic-BLUP method, where the BLUP estimation was performed using the relationship between individuals derived from molecular marker data, together with pedigree data. Genotyping data was available only for the clones present in this study and not their parents, but pedigree information was available for all individuals so the $\boldsymbol{H}$ matrix method was employed in order to combine both sources of relationship information.

3.2 Materials and methods

3.2.1 Phenotypic data

Phenotypic information used in this study was previously described in (Brum et al., 2018), and is comprised of sugarcane clones from 2 subsequent breeding cycles (*Cycle* '05 and *Cycle* '06) from the CTC - Centro de Tecnologia Canavieira breeding program, located in Piracicaba, São Paulo, Brazil. In total there were 35,284 records per trait from the two cycles (table 3.1). Each cycle had 3 stages of data; on *Stage 1*, biparental records were recorded by family, whereas *Stage 2* and *Stage 3* had records for individual clones, as they were clonally multiplied

from selections carried out after the previous cycle. In *Stages* 1 and 3 families and clones there were two replicates but *Stage 2* had a single replicate.

The traits considered for this study were Brix and plot weight, which are both available only in *Stage 1* and *Stage 3*. For this reason, those are the only stages considered in this study. Only *Stage 3* had evaluations taken for 2 years, which were the plant crop (first harvest, taken on plants grown from clonal multiplication) and ratoon crop (second harvest, taken on plants regrown from plant crop).

None of the biparental crosses were repeated between the cycles so none of the clones in *Stage 3* are in common between the cycles, but checks were included. Despite no repetition of crosses, 25% of the parents used in *Cycle* '05 were also used in crosses in *Cycle* '06.

Due to the selection that occurred within a cycle (from *Stages* 1 to *Stages* 3) the size of families present in *Stage* 3 were variable, ranging from 1 to 48 clones. This also causes the sizes of progeny from a parent to be variable in *Stage 3*, and as the number of crosses in which a parent took part in *Stage 1* was also variable, the size of progenies from a given parent was also variable since the first stage. As a consequence of the selection in early stages and the preferential use of parents in crosses, the number of phenotypic records for the progeny of a given parent varied, as shown in fig. 3.1.

Table 3.1: Field data overview. Number of genotypes, phenotypic records and related statistics are organized by their originating *cycle*, *stage* and *season* if applicable.

Cycle	Stage	Season	Records[a]	Genotypes[b]	in $\boldsymbol{K}$[c]	Families	Females	Males	Total # parents	Planting Year	Harvest Years
'05	1	-	3943	1718	0	1718	605	147	650	2005	2006
'05	2	-	9379	9209	1	1558	580	146	637	2007	2008
'05	3	early	2680	610	534	389	246	69	280	2009	2010/2011
'05	3	late	2768	629	572	443	246	94	293	2009	2010/2011

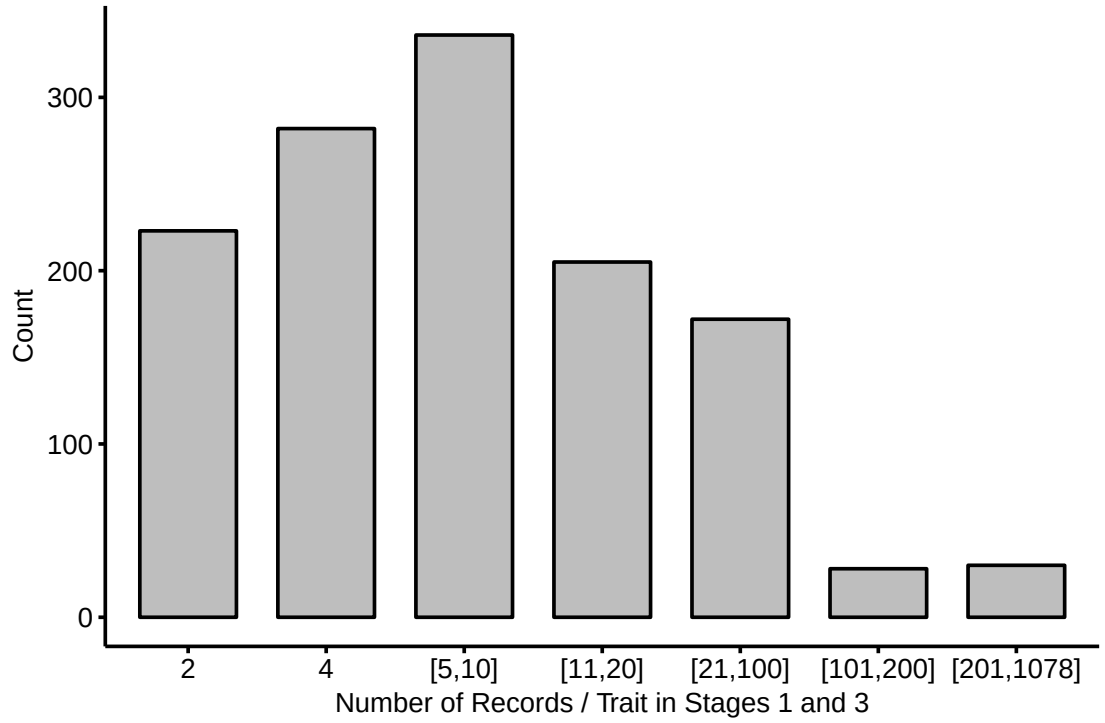

Figure 3.1: Amount of information available per parent, in terms of the number of phenotypic records of families (*Stage* 1) or clones (*Stage* 3) derived from a parent. For parents with 5 or more records, the values were grouped in closed intervals.

Cycle	Stage	Season	Records[a]	Genotypes[b]	in $\boldsymbol{K}$[c]	Families	Females	Males	Total # parents	Planting Year	Harvest Years
'06	1	-	3122	1457	0	1457	629	137	681	2006	2007
'06	2	-	9592	9187	2	1269	633	160	699	2008	2009
'06	3	early	1824	423	259	273	198	89	255	2010	2011/2012
'06	3	late	1976	449	397	285	216	88	274	2010	2011/2012
		Total	35284	16805	1755	3322	1240	249	1306		

Note: [a] Number of field records per trait, including replicates and checks. [b] Refers to families when *Stage* 1, clones when *Stage* 2 and 3. [c] Genotyped individuals, includes checks.

3.2.2 Genotypic information

Pedigree data for all crosses in both *Cycle '05* and *Cycle '06* were available and covered an average of 10 generations (Atkin et al., 2009). With that information, kinship relationship between all pairs of clones and parents were estimated using the package `pedigreemm` (Vazquez et al., 2010), deriving a relationship matrix, $\boldsymbol{A}$.

Clones from *Stage 3* in both cycles were genotyped using sequencing of DNA targeted by pre-selected DNA probes, and SNP detection. Genotyping procedure and filtering were previously described in Brum et al. (2018). In total there were 54,675 SNPs from 1,778 different clones, with a maximum of 25% missing data per clone. Markers were coded with 0 for the presence of only the reference allele in the probe sequence, 1 for both reference and alternative alleles present, and 2 for presence of only the alternative allele. A genomic relationship matrix ($\boldsymbol{K}$) was then obtained with the formula:

$$\boldsymbol{K} = \frac{\boldsymbol{M}\boldsymbol{M}^t}{k} \tag{3.1}$$

Where $\boldsymbol{M}$ is the mean centered molecular marker matrix with SNP information in columns and clones in rows. The scaling constant k was chosen to be $mean(diag(\boldsymbol{M}\boldsymbol{M}^t))$ so that the $\boldsymbol{M}$ diagonal has comparable scale to the pedigree derived relationship matrix, $\boldsymbol{A}$, which also has a mean diagonal value close to 1.

In order to obtain relationship matrices $\boldsymbol{H}$ that combine pedigree and molecular marker information, the "Whole Pedigree" formulation from Legarra et al. (2009) was used. In our case, the molecular marker matrix $\boldsymbol{K}$ has a subset of the

individuals that are present in the pedigree relationship matrix, $\boldsymbol{A}$ (expected relationships), and the resulting $\boldsymbol{H}$ will be the same size (same individuals) as $\boldsymbol{A}$. For the $\boldsymbol{G}$ matrix (Legarra et al., 2009) that informs the observed relationships in the computation of $\boldsymbol{H}$, we utilized two formulations, one where $\boldsymbol{G} = \boldsymbol{K}$, and another in which $\boldsymbol{G} = \frac{1}{2}\boldsymbol{A}_{22} + \frac{1}{2}\boldsymbol{K}$, where $\boldsymbol{A}_{22}$ is the submatrix of $\boldsymbol{A}$ containing relationships for the same individuals that $\boldsymbol{G}$ has. When this latter formulation for $\boldsymbol{G}$ is used we will refer to the resulting $\boldsymbol{H}$ matrix as $\boldsymbol{H}_w$ in order to distinguish from the other case.

3.2.3 Genomic prediction and validation

For a set of phenotypic measurements y_{ijkl}, where l refers to the l^{th} measurement (from different replicates, years, or conditions) for a given trait taken on clone k derived from a cross between parents i and j, the "clonal value" from clone k was derived from measurements given by $y_{ijk.}$, whereas the breeding value for parent i was derived from measurements $y_{i...}$ and $y_{.i..}$, referring to all phenotypic measurements taken on clones whose male or female parent was i. Settings that differentiate where i was used either as female parent or as male parent are possible too, but were not considered here due to unbalanced data (parents were more frequently used as the male parent).

Genomic prediction for parents was performed using the linear mixed model in a single step:

$$y_{ijklmnp} = \mu + \gamma : \omega_{ij} + \theta_k + a_l + b_{m[l]} + g_n + e_{ijklmnp} \tag{3.2}$$

Where $y_{ijklmnp}$ refers to a phenotypic measurement, μ: the overall mean, $\gamma : \omega_{ij}$: the interaction term for *Season* i and *Cycle* j, θ_k: *Harvest* k, a_l: *Field* l, $b_{m[l]}$: *Set* m nested within *Field* l, g_n: genotype n, and e_{ijklmn}: residual error. Both μ, $\gamma : \omega_{ij}$ and θ_k are considered fixed effects. The effects a_l, $b_{m[l]}$, $e_{ijklmnp}$ are independent and normally distributed. The phenotypic measurement values for $y_{ijklmnp}$ were taken from clones from *Stage* 3 and families from *Stage* 1, with their respective replicates and spanning the different seasons, cycles and harvests. Because not only *Stage* 3 but also *Stage* 1 was used, Brix and Plot weight were the only traits considered. The genotype effect is normally distributed as $g_n \sim \mathcal{N}\left(0, \sigma_g^2 \boldsymbol{H}\right)$, with σ_g^2, the genetic variance component. The covariance matrix $\boldsymbol{H}$ is the additive covariance between each genotype n, and encompassed all individuals that were phenotyped directly and also the parents used in the families evaluated in *Stage* 1.

By fitting eq. 3.2 using either matrices $\boldsymbol{A}$, $\boldsymbol{H}$, or $\boldsymbol{H}_w$, BLUP estimates for all parents in the matrix can be obtained (Robinson, 1991; Piepho et al., 2008). Three scenarios were used for fitting eq. 3.2: 1) only *Cycle* '05 data included, 2) only *Cycle* '06 data included, and 3) both *Cycle* '05 and *Cycle* '06 data included in the model fitting.

The training set (for a prediction/validation procedure) consisted of the subset of those parents for which there were always phenotyped progeny in the fitted data (in $y_{ijklmnp}$). Whereas the validation set consisted of another set of parents that still have breeding values estimated, but in the prediction step their progeny were not included in the fitted data, and on the estimation step their progeny was included, thus allowing the comparison between predicted and observed breeding values. Parents whose progeny were exclusively evaluated

in *Cycle* '06 were then considered the validation set, and used to assess prediction accuracy. Therefore scenarios 2 and 3 provided breeding value estimations to *Cycle* '06 parents, as it was based on observed data of their progeny, whereas scenario 1 provided only predicted breeding values for *Cycle* '06 parents, as none of their progeny were evaluated in *Cycle* '05.

In order to estimate prediction accuracy, let x be the breeding values of the *Cycle* 6 exclusive parents, estimated under scenario 1 (training) using all relationship matrices, and let y be the breeding values of these same parents estimated either under scenarios 2 or 3 (validation), using only the matrix $\boldsymbol{A}$. Then the correlation between x and y was used as the estimate of prediction accuracy for those different conditions. Model fit was performed in R and mixed models program `wombat` (Meyer, 2007).

3.3 Results

3.3.1 Prediction Accuracies

Parental prediction results are shown in fig. 3.2, grouped by trait. When using only data from *Cycle* '06 to estimate breeding values from parents in the validation set, the correlations with breeding values estimated using data from *Cycle* '05 varied from 0.13 to 0.14 for Brix, and from 0.20 to 0.25 for plot weight. The best estimate for Brix was obtained when predictions used molecular marker derived relationships and the matrix $\boldsymbol{H}_w$ provided the best result. For plot weight the use of only pedigree data provided the best predictions.

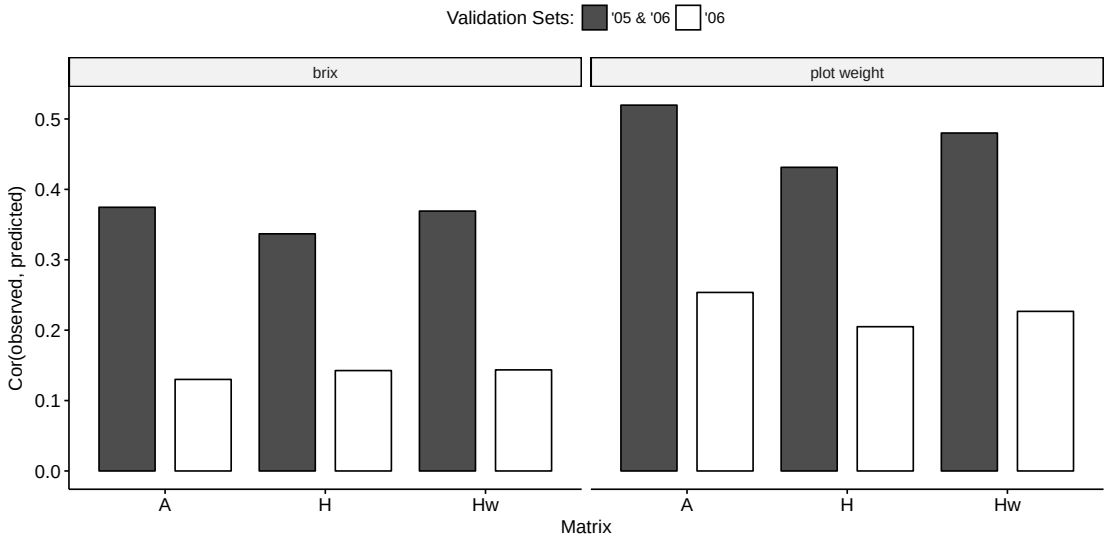

Figure 3.2: Prediction accuracy for breeding values of the traits Brix and plot weight of parents used in the breeding program. Training set has data only from *Cycle* '05, whereas the breeding values for parents exclusively used in *Cycle* '06 (Validation Set) are estimated either using '06 data or using both '05 and '06 data. Prediction was performed using either pedigree relationships (matrix $\boldsymbol{A}$) or a combination of pedigree relationships and molecular marker derived relationships (matrices $\boldsymbol{H}$ and $\boldsymbol{H}_w$).

We also observed an increase in accuracy of estimates when data from *Cycle* '05 was included to support the estimation of breeding values of the validation set (fig. 3.2), with estimates varying from 0.34 to 0.37 for Brix, and from 0.43 to 0.52 for plot weight. Similar to the previous case, the use of matrix $\boldsymbol{H}_w$ provided the best prediction for the trait Brix, and the matrix $\boldsymbol{A}$ the best prediction for plot weight.

Among the 623 parents present exclusively in *Cycle* '06, that were then used as a validation set in previous results, there were 118 parents that belonged to 22 full sib families (average of 5.36 genotypes per full sib family). Using

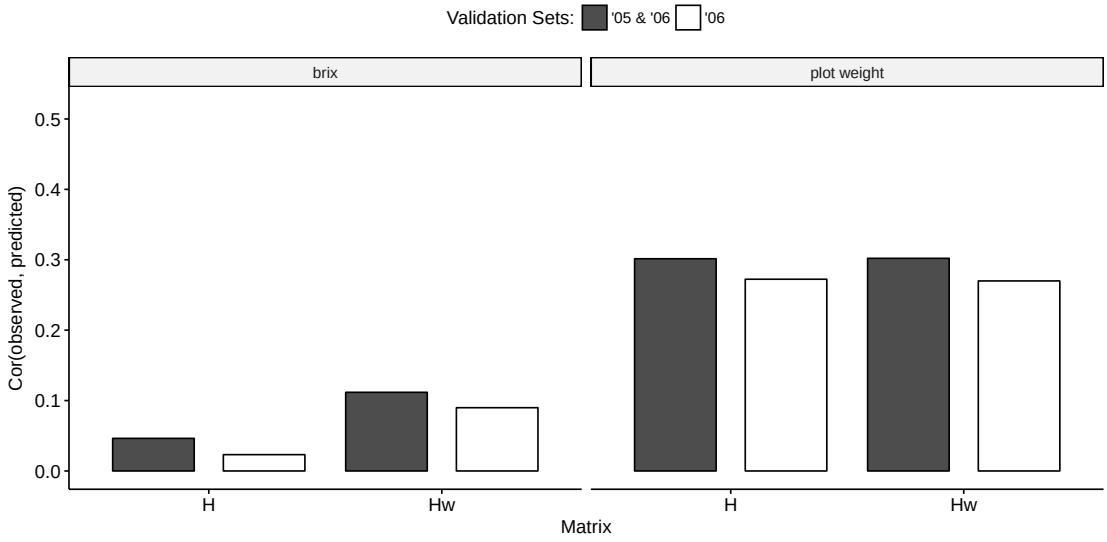

Figure 3.3: Prediction accuracy for parents grouped in full sib families. Bars are average of accuracies estimated for each family. Training set has data only from *Cycle* '05, whereas the breeding values for parents exclusively used in *Cycle* '06 (validation Set) are estimated either using only '06 data or using both '05 and '06 data. Prediction was performed using a combination of pedigree relationships and molecular marker derived relationships (matrices $\boldsymbol{H}$ and $\boldsymbol{H}_w$).

only those parents, we estimated prediction accuracy within full sib families of parents (fig. 3.3). As parents in the same family have the identical pedigree, their prediction through matrix $\boldsymbol{A}$ renders the same value, and can not be used for estimating correlations. Therefore fig. 3.3 only presents results for matrices $\boldsymbol{H}$ and $\boldsymbol{H}_w$. Prediction accuracies varied from 0.02 to 0.11 for Brix, and from 0.27 to 0.30 for plot weight. In comparison to previous results, there was a reduction of prediction accuracies for all cases, and also a reduction in the difference between the validation set that uses only *Cycle* '05 data for estimation and the one that uses both *Cycles* '05 and '06.

3.4 Discussion

In this study, genomic prediction for an applied sugarcane breeding program was evaluated for prediction of breeding values for parents. It is common to focus only on selecting clones and then use them as parents. On the other hand, it is necessary to properly evaluate the merit of sugarcane parents in crosses and their potential to produce varieties. Traditionally, parents are selected only after their performance as clones has been determined so the parent selection is delayed for several years. This makes the phenotypic traits that inform this evaluation more valuable but it extends the length of breeding cycles as measured from the time when the cross that produces a future parent is performed and adequate information about the future parent's performance is available.

3.4.1 Prediction of parental performance

We performed predictions focusing on parental breeding value prediction taking into account additive genetic effects to predict their breeding values. Those additive effects were estimated exclusively from observed data on the traits Brix and plot weight from the progeny of the parents, not from observed data from the parents themselves.

The use of progeny data provided opportunities and limitations. First of all, it was possible to estimate breeding values for the parents used for crosses in *Stage* 1, which are also the parents of clones in *Stage* 3. One advantage is the possibility of using both *Stage* 3 clonal and *Stage* 1 family data. It is common in sugarcane breeding programs to use *Stage* 1 family data for breeding value

estimation. It reduces costs by discarding poor parents that will not be further evaluated through progeny in downstream stages. But the disadvantage of this approach is that performance as clones and performance in crosses can differ. The use of both *Stage* 3 clonal and *Stage* 1 family data might address this disadvantage, and it provides a larger dataset for the value estimations that were from multiple years. Genotype by environment interaction (GxE) is an important factor to be taken into account when interpreting the prediction accuracies. The training set used data from both *Stages* 1 and 3 allowing the estimation of breeding values from multiple-year data, for the particular years that composed this training data. At the same time, the validation set was composed with data from different years as well and might have impacted the prediction accuracies.

Another consideration is the unbalanced data caused by the different amounts of progeny data available for the parents (fig. 3.1). The unbalanced data will affect the reliability of the breeding value estimations, which can be offset by the number and relatedness between parents in the relationship matrices.

The prediction accuracy in breeding value estimations were correlated with estimated values derived either from *Cycle* '06, or from both *Cycle* '05 and '06. Performances of predictions from both cycles were higher as expected from the fact that relationships among parents present in both cycles will cause the estimates to be influenced by the phenotypic records from the first cycle. On the other hand, those are the best estimates possible for the parents when predicting their use in future crosses. As more phenotypic data from similar conditions are accumulated, (location, breeding population, traits considered) we can expect prediction accuracies to be intermediate between those from when only *Cycle*

'06 was used as validation and those when both cycles were used.

Prediction accuracies of parental breeding values were higher when using markers for Brix but not for plot weight (fig. 3.2). In cases where pedigree records are not available or for prediction within full sibs, the molecular markers have a clear advantage. This was evident, despite the smaller dataset, when prediction for full sib parents was assessed (fig. 3.3).

The predictions for parental breeding value used relationships derived from a combination of pedigree and molecular markers in the $\boldsymbol{H}$ matrix derived from all parents, *Stage* 1 families and *Stage* 3 clones. So one venue for further research would be to use a dataset where the parents themselves were genotyped. Parental breeding value prediction accuracies for full sib parents (fig. 3.3) were comparable in magnitude to using all parents (fig. 3.2). Prediction accuracies using the $\boldsymbol{H}$ matrix may be similar to those using only markers.

3.5 Conclusions

Genomic Selection can be a tool for improving the efficiency of a sugarcane breeding program through its use in parent selection. Previous studies on GS in sugarcane (Gouy et al., 2013; Brum et al., 2018) validated their predictions against total genetic values of clones, or clonal values pointing to the use of GS to improve the clonal selection in a breeding program. In that situation, efficiency gains come from the selection process, and stages could be skipped, saving time in the selection stages. But the clonal propagation rate might limit the applicability of this strategy because moving from an initial stage with little plant material, to the next stage, where abundant plant material is required to

plant large plots, or plots in multiple locations, is limited.

By using parental GS selection, instead of skipping stages, the cycle itself can be reduced, the length in time required by a new clone to be ready for crossing being the limiting factor. Possible approaches for using parental GS in a breeding program would be:

- Instead of phenotyping all new parent candidates, from seedling they would go directly to crossing.
- Focus phenotyping (progeny testing) on the most promising candidates (according to predictions), for instance, using multiple location testing on them.
- Use recurrent selection for the traits with higher accuracy with cycles reduced to the minimum for first flowering (1, or 2 years).
- Selection within families, as it would not be allowed by pedigree-BLUP prediction, and the most challenging phenotyping would occur within families, or the most promising selection of new parents would occur within families.

CHAPTER 4

ON THE USE OF MOLECULAR MARKER PHENOTYPES IN A POLYPLOID GENOTYPE

4.1 Introduction

Sugarcane is an autopolyploid with an unusually complex genetic configuration due to its high ploidy level with values between 8 and 12, presence of aneuploidy and a hybrid genome with most of its content originating from *S. officinarum* L. and with a minor proportion from *S. spontaneum* L. (Piperidis et al., 2010).

The complex configurations, specifically the multiple dosage of alleles, is difficult to observe directly in molecular marker phenotypes such as Restriction fragment length polymorphism (RFLP), Diversity Arrays Technology (DArT) and simple sequence repeats (SSR) markers, which are traditionally used in sugarcane research. Most genetic mapping studies used single dose markers and filtered out higher dosage markers in the first step for linkage mapping (Wu et al., 1992; Ripol et al., 1999; Aitken et al., 2007), or used markers with low dosage (Garcia et al., 2006).

In other polyploids, markers are generally better understood and their dosage is used in genetic studies. For allopolyploids, alleles segregate in a diploid fashion leading to a diploid analysis of the molecular marker phenotypes. There are also the cases of lower ploidy level (tetraploidy to hexaploidy) where the identification of dosage is feasible.

So is it possible to do the same in sugarcane? The use of higher dosage (2 and higher) for molecular markers in sugarcane has been discussed, in some references: Cordeiro et al. (2006) demonstrated the identification of SNP molecular markers in sugarcane, using pyrophosphate sequencing, and discussed the use of SNP base frequencies to determine the likely number of alleles. Garcia et al.

(2013) demonstrated the use of statistical clustering techniques for dosage and ploidy level estimation in sugarcane.

These previous studies are based on the use of ratios of signals specifically derived from the different alleles as the phenotype of molecular markers in sugarcane. In order to determine the dosage of corresponding alleles in sugarcane, a number of conditions are required: 1) ploidy level at the locus is known or can be determined 2) signal is derived from every one of the alleles present at a locus under consideration 3) signals are derived only from the target locus 4) signal intensity is quantitative in relation to the dosage of alleles.

But those assumptions hardly can be met for sugarcane, and possibly even for other high level autopolyploids more well behaved. The problems are: Aneuploidy is present in sugarcane, leading to lack of *a priori* knowledge of the ploidy level. It is even possible that aneuploidy will cause ploidy level differences between different chromosomes, for a given individual genotype. Determination of ploidy level directly from molecular marker data in sugarcane has been demonstrated in Baker et al. (2010) and Garcia et al. (2013). Nevertheless, for Baker et al. (2010) the use of genotypic data from progeny of biparental crosses was required and this limits the application of the methodology (as for GWAS and GP). Statistical power was limited since it was difficult to keep large sugarcane biparental populations. Selection can an be an issue, leading to distortions in the segregation frequencies. The results will only apply to the ploidy level of parents used in the cross, not their progeny, which is of greater interest when mapping QTL. In the case of Garcia et al. (2013), the experimental results presented in the paper show a wide range of ploidy levels (from 6 to 20) for a given genotype, but this was not confirmed from FISH/GISH results shown

in the paper. The results from statistical analysis need to be verified. It is also possible that some of the inconsistencies observed in Garcia et al. (2013) are due to issues related to assumptions as discussed below.

In the case of the second condition, sugarcane is highly heterozygous (Vettore et al., 2003, Garsmeur2011) and because of polymorphisms near a target locus, the signal being measured might not be derived from every allele present. In the case of molecular marker detection methods that use PCR (such as Sequenom), primers that are used for DNA amplification may not hybridize to every allele at a locus. This will be a common theme for all hybridization-based methods (as in Illumina chips). For GBS and other highly multiplexed sequencing based methods (RapidGenomics' CaptureSeq, e.g.), stringency in parameters for read clustering, combined with the presence of polymorphism may cause reads from the same locus to be grouped in different clusters, or not clustered at all. This can result in not detecting some alleles, thus skewing the signal ratio between the alleles that were detected.

For condition 3, the opposite of 2 can also happen when the signal derived from more than one locus is detected together as one marker. It can arise from duplication of genes being targeted, or repetitive regions, which are common features of plant genomes.

Both conditions 2 and 3 could be better addressed if the whole genome sequence of sugarcane were known. Efforts for this sequencing are underway (Okura et al., 2016). The quality and how representative the end result is of the sugarcane germplasm in use is important for the ability of addressing conditions 2 and 3.

For condition 4, even with exclusive observation of all alleles from a locus, the observed signal might not be accurate, due to signal error detection at the equipment level, or measurement error. For instance, in a tetraploid with only two alleles at a locus, and an allele present in single dose, one would expect an allele dosage ratio of 1:3. If using a sequencing based method for genotyping, and one obtains 10 reads from this locus, we would expect between 1 and 3 reads for the single dose allele, and the remaining to the second allele. Nevertheless, due to sampling variation, a 1:1 ratio (4, 5 and 6 reads), for instance, would happen with probability 22.1%, causing the observer to believe that both alleles are present in double dose.

We are proposing then, to use a method that relies in fewer assumptions. i.e. to use directly the ratio of the allelic signals, without dosage (and ploidy level) determination. It avoids condition 1, and reduces the importance of conditions 2 and 3, for some applications of molecular marker data (prediction). Condition 4 is not relaxed, but its effect can be reduced with replication of data (for Sequenom, Illumina chips), or increase in signal quality (e.g. in case of sequencing, requiring high read depth).

Using this tool, we sought to address the hypothesis, that the use of higher dosage marker phenotypes is an advantage for genetic studies in sugarcane. We will test this in the context of association analysis (GWAS) and Genomic Prediction (GP), comparing results that can be obtained using the same set of molecular markers interpreted in two different ways: the ratio of allelic signals as mentioned above, and simply presence or absence of alleles. In the former case, the high dosage of alleles is expressed in the molecular marker information, in contrast to the latter case. In this study we evaluate the effect of use

of dosage information, measured in terms of its value for a molecular marker phenotype, for genetics studies in sugarcane.

4.2 Methods

4.2.1 Molecular data

The initial molecular marker dataset obtained from sequencing, had 245,923 SNPs from 1,904 samples. All SNPs were filtered to be bi-allelic in the genotype-calling stage, and read counts for each of the two alleles, the reference and the alternative allele, were available which in combination provide a read depth count per datapoint (marker per sample). In the initial marker dataset, 46 samples had more than 50% missing data and were removed. The initial marker set had a maximum 34% missing data points. But 148 markers had an outlier number of read counts and were excluded as well.

SNP coding

We then adopted two systems for coding the SNP calls: the *discrete* coding, in which the genotypes homozygous for the reference allele, heterozygous (presence of both alleles) and homozygous for the alternative allele were coded as 0, 1 and 2, respectively; and the *continuous* coding where the ratio between the read counts for the reference allele and the total number of read counts was used as the genotype. Therefore the genotype of a datapoint (a molecular marker value for a given individual DNA sample) is a continuous value between 0 and 1. This

is exemplified in table 4.1.

Table 4.1: Examples of genotype coding using *discrete coding* and *continuous coding*. Each Datapoint is a different marker and genotype combination.

Datapoint#	Reference Allele	Alternative Allele	Read Count Reference	Read Count Alternative	Discrete Genotype	Continuous Genotype
1	A	T	10	0	0	1.00
2	A	T	12	8	1	0.60
3	C	G	2	10	1	0.17
4	T	G	0	8	2	0.00

The use of the *continuous* coding may provide access to more information to differentiate individuals from the molecular marker data. This is exemplified in the fig. 4.1, where the distribution of genotypic values for two different markers is shown in Panels **A** and **B**, under the *discrete* and the *continuous* coding for the same datapoints. In Panel **A**, the molecular marker in the *discrete* coding (bottom plot) has a larger number of individuals under category "1", which represents the heterozygous state (both alleles of the SNP are present), but all these individuals are spread in values from 0 to close to 1 if *continuous* coding is used (top plot). The Panel **B** exemplifies a case of a molecular marker that shows only the heterozygous state in *discrete* coding (bottom plot). Such a SNP would not be useful to distinguish individuals if *discrete* coding was used, but as the top plot in fig. 4.1 shows, the use of *continuous* coding enables the use of the marker.

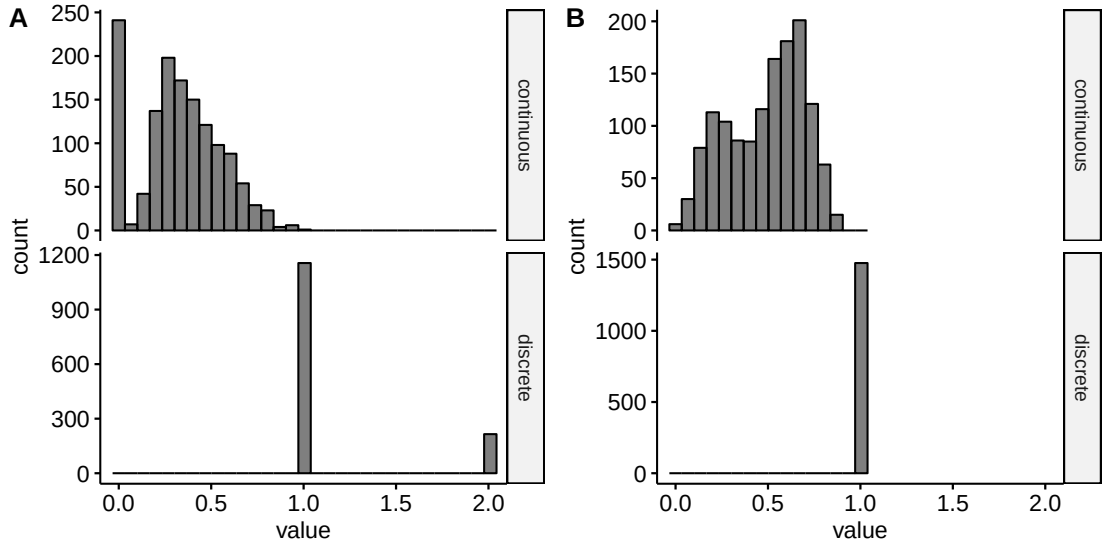

Figure 4.1: Histograms of the distribution of values in the *discrete* and *continuous* coding across all the genotyped individuals. Panels **A** and **B** show examples of different sugarcane molecular markers from the dataset in use here. The same marker is shown under *continuous* coding (top) and *discrete* coding (bottom)

SNP filtering on read depth

The read depth (the total number of reads for all alleles) may influence the quality of the genotypic value, independently of the coding. For the *discrete* coding, a low read depth may lead to the non observation of a given allele, causing, e.g. a heterozygous genotype to be observed as homozygous for the only observed allele. In the case of the *continuous* coding not only the absence of an allele is a possibility, but also a lower read depth may lead to larger variance in the observed ratio genotype. In order to check which read depth would be ideal for filtering the data, 20 samples that had been genotyped two or three times in the dataset were analyzed. Previously we have used error-rate in duplicated samples to define an optimal read depth requirement. The error-rate statistic is the ratio of datapoints that have different values to the total number of dat-

apoints, given that the datapoints came from two distinct DNA samples from the same plant or clone. In the case of *continuous* coding, calling two datapoints to be equal or different may require the definition of a threshold of similarity, that should be determined independently. An alternative statistic would be the use of variance per locus. One would expect that the variance at a locus would be zero for replicates of the same DNA. But the variance may correlate with allele frequency or heterozygosity of a given molecular marker. In order to avoid this issue, we made use of two opposing statistics, the Euclidean distance taken from molecular marker values between replicated samples, and the Euclidean distance from non-replicated samples. For a varying threshold of read-depth, fig. 4.2 shows the mean distance computed only between replicated genotypes and mean distance only among different genotypes (left panel), and the ratio between these two distances (right panel). Based on this ratio, the best read-depth is 80, with a ratio of 0.56 between the distances. The filtering based on read depth will introduce missing datapoints, so that by requiring a minimum of 50% non missing data per genotype after requiring minimum 80 read-depth, we will end up with only 13,379 molecular markers. Figure 4.3 shows the reduction in number of molecular markers and samples as it is required at least 50% non-missing data for both markers and samples.

Taking into account the variance in the distances computed for the different pairs of genotypes, we observe that there is non-significant difference between the ratio of distances at read depth 80 and 50, whereas this last read depth provides a larger amount of markers, 48,780. The read depth of 50 was used then as the threshold for filtering data points, with further requirement that both for markers and samples the amount of missing data is not larger than 50%. These criteria resulted in a final data set of 48,830 SNPs in 1,476 different genotypes.

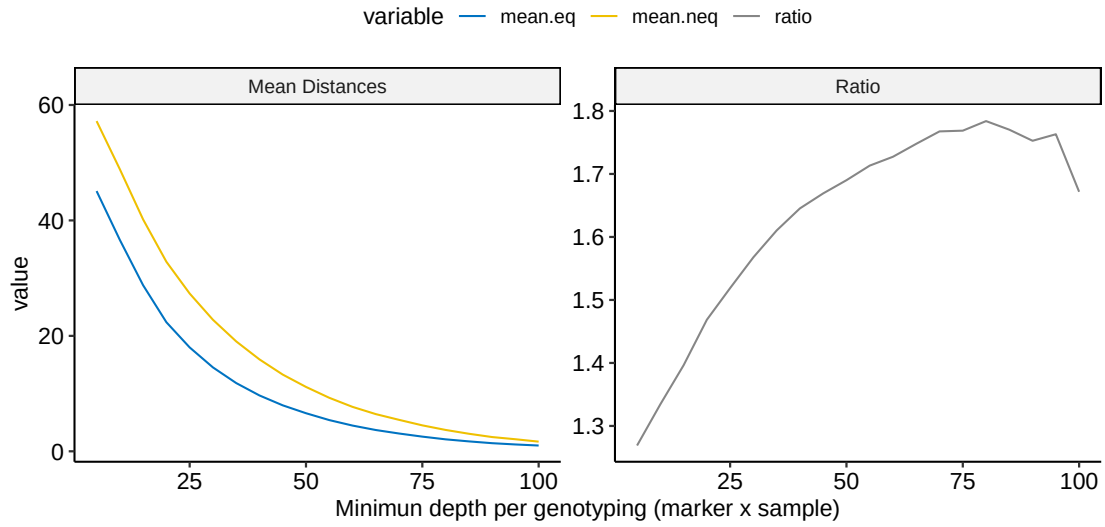

Figure 4.2: The mean Euclidean distance was computed between pairs of genotypes filtering datapoints on different thresholds for read depth. On the left, the mean distance for pairs of replicated genotypes (`mean.eq`) and pairs of non-replicated genotypes (`mean.noneq`) is presented. On the right, the ratio (mean distance of non-replicated over replicated pairs) between these distances is shown.

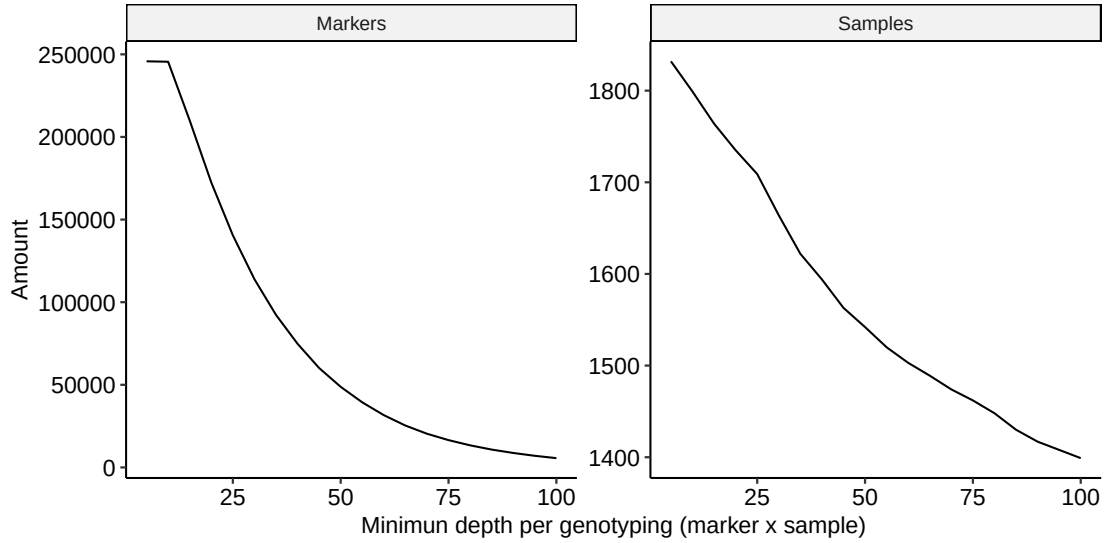

Figure 4.3: Reduction in number of samples and markers as the requirement of read depth increases.

Molecular marker matrices formulation

Among the 48,830 SNPs in the final set, 6,706 were similar to the one presented in fig. 4.1, panel **B**, where most of the genotype calls were heterozygous if the *discrete* coding is used, but there was more genotypic variation in *continuous* coding. Here a threshold of 1% of non-heterozygous genotype calls was used for the *discrete* coding version, and a threshold of 0.05 standard deviation for the *continuous* coding, which we will refer to as "polymorphic" markers, in the sense that they are polymorphic in both coding systems.

With this result, 3 versions of the molecular marker matrix were obtained:

- Matrix $\boldsymbol{D}$: using *discrete* coding with only the 42,124 markers that were polymorphic.
- Matrix $\boldsymbol{C}$: using *continuous* coding with only the 42,124 markers that were polymorphic.
- Matrix $\boldsymbol{C_2}$: using *continuous* coding with all the 6,706 markers not included in the previous matrices.

For each of the above matrices, a corresponding relationship matrix among the genotyped individuals was estimated using equation 4.1:

$$\boldsymbol{K} = \frac{\boldsymbol{M}\boldsymbol{M}^t}{k} \tag{4.1}$$

Where $\boldsymbol{M}$ is the mean imputed and centered molecular marker matrix with SNP information in columns and individuals in rows, obtained from $\boldsymbol{D}$, $\boldsymbol{C}$ and $\boldsymbol{C_2}$. The scaling constant k was chosen to be $mean(diagonal(\boldsymbol{M}\boldsymbol{M}^t))$.

How these different representations of molecular marker data impacts the overall relationship between pairs of individuals was analyzed through Principal Component Analysis (PCA), computed by the Singular Value Decomposition of the relationship matrices from $\boldsymbol{D}$, $\boldsymbol{C}$ and $\boldsymbol{C_2}$.

4.2.2 Association Analysis

In order to assess the effect of the coding system for markers in the ability to study quantitative traits, we performed association tests for all markers. For the sugarcane population under consideration, phenotypic records for two breeding cycles were available, Cycle '05 and '06, and the former was used for the association analysis as it contained a larger number of individuals. The phenotypes Weight (total biomass weight from harvested plot), Brix (total soluble solids %, mostly sugars), Pol (sucrose content in cane juice as % of biomass, measured by polarimetry) and Fiber (non soluble solid content from stalks, in % weight) were measured for plant crop (first harvest) and ratoon crop (second harvest), with measurements taken in Early or Late season genotypes. Due to the higher broad sense heritability of Fiber, it was used in the results reported here. In total there were measurements from 484 clones taken in the early season, and 424 clones in the late season, with these same clones being harvested for the plant crop and ratton crop. Those four conditions were considered as four traits in the association tests.

In order to allow non-additive relationships between marker values in the *continuous* coding and phenotypes, a Generalized Additive Model (GAM) was fit to marker effects, as well as a regular additive linear model (LM). The GAM

model was not fit to the *discrete* coding system.

As shown in fig. 4.4, population structure was observed in this population due to more frequent use of some parents in the crosses that formed the population. In order to control for population structure in association tests, the pedigree and four principle components of the population were used.

Quantile-Quantile plots for the p-values of the association tests were used to assess the effect of population structure on the statistical significance of tests. The EMMAX method (to use a random effect with covariance from pedigree matrix or marker matrix) or principal components from marker matrix were evaluated. The control of population structure using pedigree data was then preferred due to its independence from the molecular marker values. To visualize regions associated to the traits, Manhattan plots were used, where the marker p-values were plotted at the genomic position of the corresponding marker. This position was obtained from alignment of the sequence of the molecular markers to the sequence from the sorghum genome.

4.2.3 Genomic Prediction

In order to further evaluate the usefulness in the *continuous* coding to study quantitative traits, we performed genomic prediction using the same population and traits used in the association analysis. Besides using genomic best linear unbiased predictor (GBLUP) for prediction, we sought to apply methods that might explore non-linear relationships between predictors in the data, including Support Vector Machines (SVM), Random Forests (RF), and Reproducing Kernel Hilbert Spaces (RKHS).

Optimization of hyper-parameters

The methods SVM, RF and RKHS rely on hyper-parameters that are not estimated through the fitting process, rather they need to be provided or estimated from the data. To do this, we made use of the phenotype records and genotypes from Cycle '06, making the hyper-parameter estimation independent from the prediction evaluation that we perform using only the Cycle '05 data.

As in the association analysis, the trait Fiber was used due to its higher broad sense heritability. In total there were measurements from 306 clones taken in early season, and 339 clones in late season for the Cycle '06, with these same clones being harvested twice. We then considered those four conditions as four traits.

For each of the methods SVM, RF and RKHS, and a range of corresponding hyper-parameter values, a cross-validation schema was performed, using five folds with three replicates. At each iteration, we recorded the correlation between the observed phenotypic values of the validation set and their predicted values based on the training population for the method under evaluation. This process was repeated for each of the four traits (Fiber trait conditions) and matrices $\boldsymbol{D}$, $\boldsymbol{C}$ and $\boldsymbol{C_2}$. The hyper-parameter values that provided the best average correlation across folds and replicates were chosen, for each prediction method (SVM, RF, RKHS), trait and matrix, and then used in the subsequent analyses.

Evaluation of Genomic Prediction

For the actual evaluation of genomic prediction, data from Cycle '05 was used since it had a larger number of individuals: 484 clones in early season and 424

clones in late season. Eight fold cross-validation with five replicates was used to evaluate the prediction accuracy for the four fiber traits. Prediction accuracy was estimated using the mean correlation between predicted and observed values for the validation set, across all iterations. The methods GBLUP, SVM, RF and RHKS were evaluated.

Prediction accuracies were obtained using the matrices $\boldsymbol{D}$, $\boldsymbol{C}$ and $\boldsymbol{C_2}$. For the GBLUP method, a second analysis was performed using a multiple-kernel model, where besides a random effect with covariance matrix from matrix $\boldsymbol{D}$, a second random effect was included with effects correlated accordingly to either matrix $\boldsymbol{C}$ or matrix $\boldsymbol{C_2}$.

4.3 Results

4.3.1 Representation of relationships in different coding systems

We observed differences, due to marker coding system, in the relationship between the genotypes in a PCA. Plots of the first two components are presented in fig. 4.4. The initial components explained a small amount of the total variance, 15.9%, 17.3% and 14.4% for the first component of respectively $\boldsymbol{D}$, $\boldsymbol{C}$ and $\boldsymbol{C_2}$. Comparing $\boldsymbol{D}$ and $\boldsymbol{C}$, we noticed that their first component correlation was high, 0.96, as well their second component, 0.95, and the correlation between all off diagonal elements in these matrices was 0.92. In the comparison between matrices $\boldsymbol{D}$ and $\boldsymbol{C_2}$, their first components had a correlation of 0.87, and their

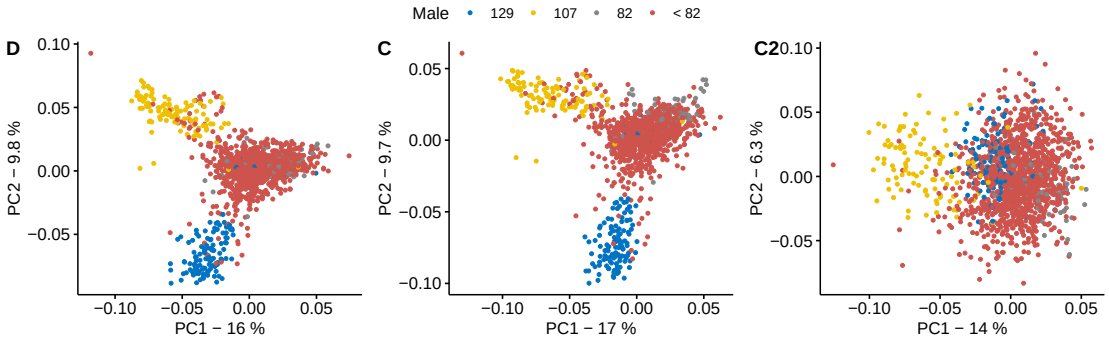

Figure 4.4: PCA plot from the first and second components obtained by singular value decomposition of the molecular marker matrices $\boldsymbol{D}$, $\boldsymbol{C}$ and $\boldsymbol{C_2}$. Each dot represents one genotype, with colors being consistent across the panels and representing the progeny of the three most frequently used male parents in this population. The amount of variation (% of the total variance) explained by the components is shown in the labels for the axes.

second components were not correlated, -0.05. Overall the correlation between off-diagonal elements in the matrices was 0.69, which is smaller than the previous comparison. The second component from $\boldsymbol{C_2}$ seemed to encode information not present in $\boldsymbol{D}$, and consequently, also not in $\boldsymbol{C}$. In fact, checking all components from $\boldsymbol{D}$, the one most similar to the second component from $\boldsymbol{C_2}$ had a correlation of 0.34.

Therefore it seems that the change in coding from *discrete* to *continuous* does not change significantly the estimation of relationships between genotypes, but it can give access to markers not polymorphic in *discrete* coding.

It remains to be checked what impact this extra information can have when analyzing quantitative traits. This is presented in the next sections.

4.3.2 Association analysis

The Quantile-Quantile plot for the p-values of the association tests is shown in fig. 4.5. There seemed to remain some effect of population structure due to the slight inflation of observed p-values in the lower end (values below 2 in the $-log_{10}$ scale), but this effect was not reduced in tests with different formulations to control for population structure. In fig. 4.5 results for the *discrete* coding are shown in the top panel and results for the *continuous* coding in the lower one, where both LM and GAM results are shown. The GAM modeling did not yield stronger associations to the traits than LM, and the lower curve of dots for GAM can be interpreted as lower power to detect associations for these cases. For this reason GAM results were not used in the Manhattam plots (fig. 4.6).

Manhattan plots showing association tests across the genome for the four Fiber traits are presented in fig. 4.6. The peaks on the y-axis have different color depending on whether the corresponding marker was tested using *discrete* or *continuous* coding. In fig. 4.6 markers from matrices $\boldsymbol{C}$ and $\boldsymbol{C_2}$ were plotted together to simplify the visualization, but it implies that there are more markers in use for the *continuous* coding than the number for the *discrete* version, increasing the chance for finding hits in the former case. Nevertheless, the visualization of results in figs. 4.5 and 4.6 does not support a clear difference between results from either of the codings.

Alternatively we computed the number of significant hits, for a range of significance thresholds from 0.001 to 10^{-5}. Results are presented in fig. 4.7, where in panel **A** matrices $\boldsymbol{C}$ and $\boldsymbol{C_2}$ were used together (as in fig. 4.6), but in panel **B** only matrix $\boldsymbol{C}$ is used. In all cases, only model LM was used. These results show that for any of the significance levels the number of significant associations was

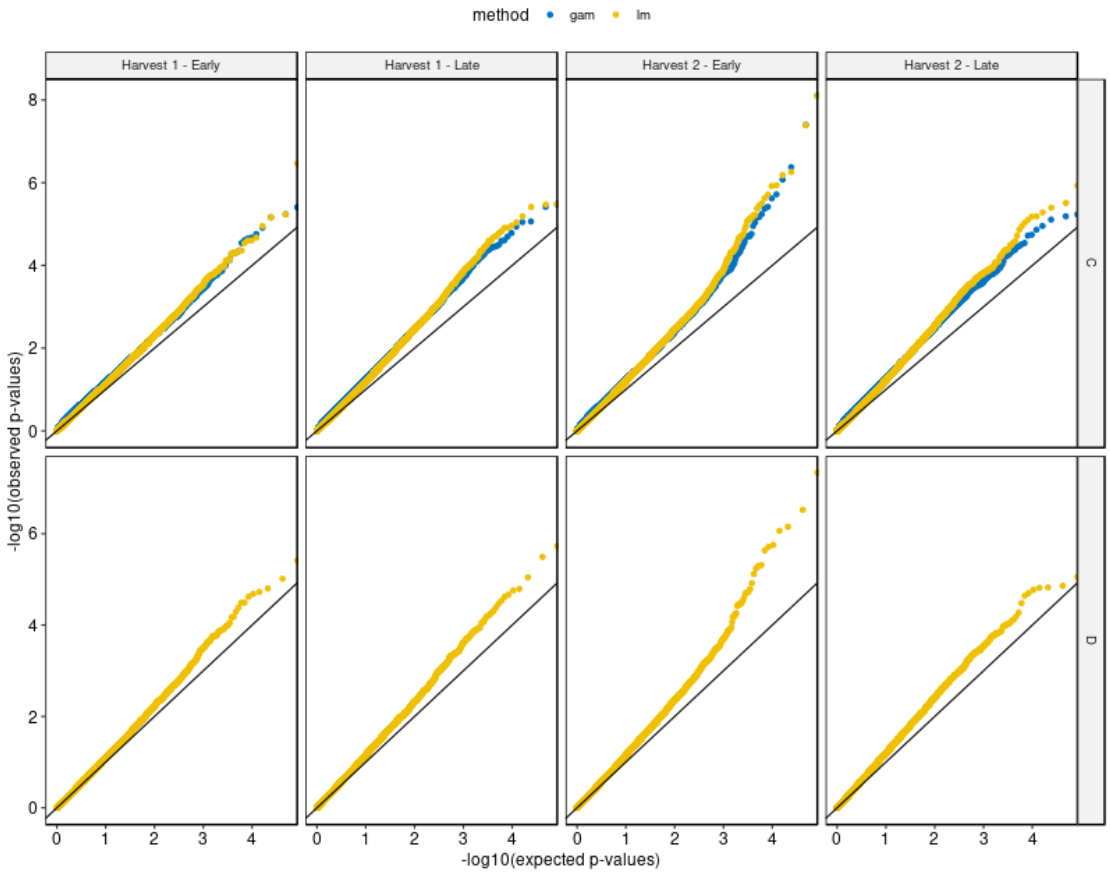

Figure 4.5: Quantile-quantile plot, comparing the expected distribution of p-values under the Null Hypothesis of no marker effect, and the observed distribution of p-values from association tests. Association tests were performed with markers in *continuous* coding (top) and *discrete* coding (bottom), for 4 fiber traits.

always higher for the *continuous* coding.

In order to compare the LM and GAM models in *continuous* coding, for the same range of significance thresholds as in fig. 4.7, we computed the number of significant markers for GAM model, that were not significant for LM modeling (table 4.2). Figure 4.8 shows three examples of significant markers (p-value < 10^{-4}) for the GAM model which were not relatively significant for LM model (p-value > 0.05), a scatter-plot of values of the markers and the values from the

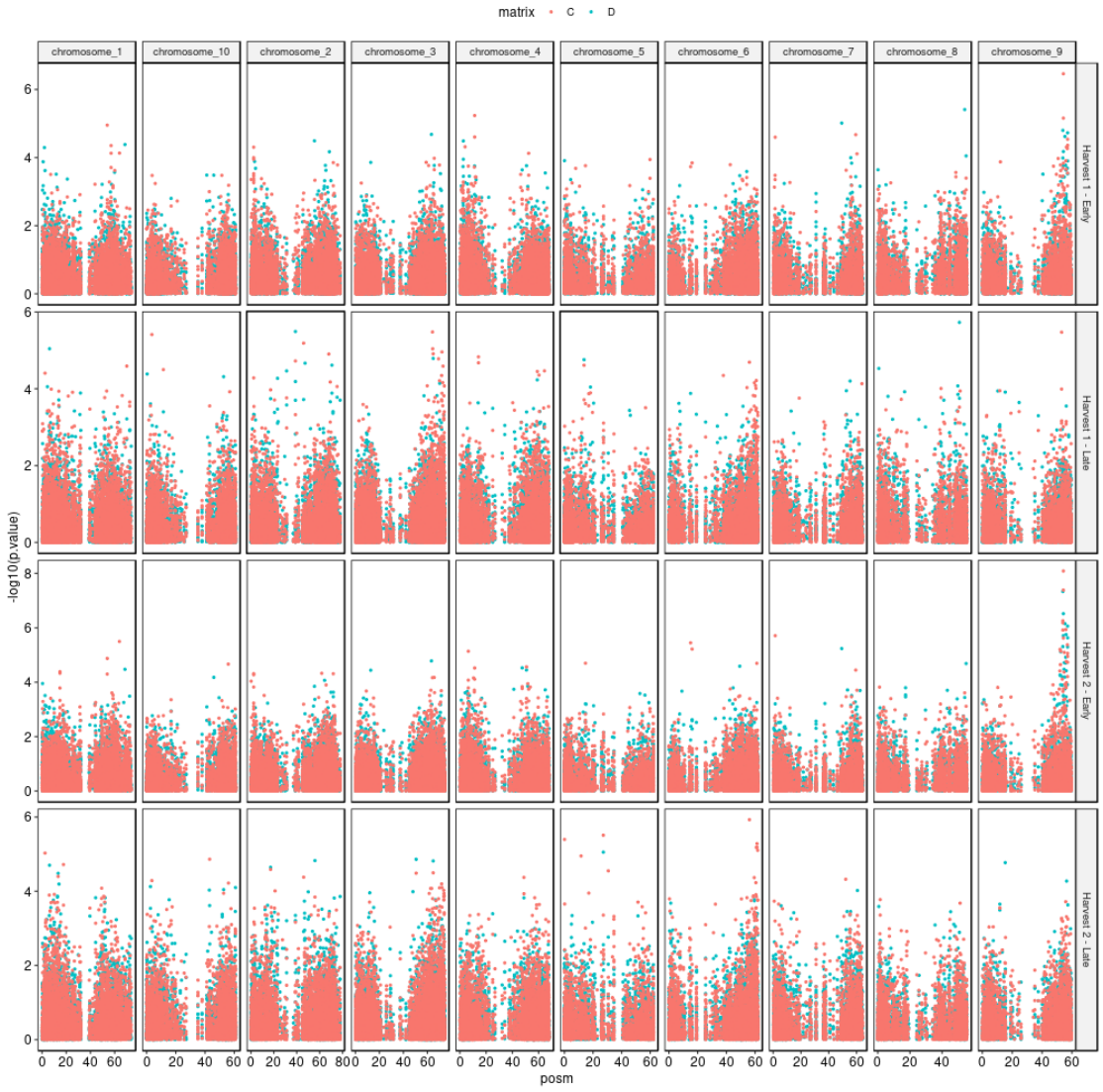

Figure 4.6: Plotting of the p-values for the association tests of molecular markers for 4 fiber traits. The markers are positioned on the sorghum chromosomes according to sequence alignment. Association tests were performed with markers in *discrete* coding ($\boldsymbol{D}$ matrix) and *continuous* coding ($\boldsymbol{C}$ and $\boldsymbol{C_2}$ matrices combined here), presented in different colors.

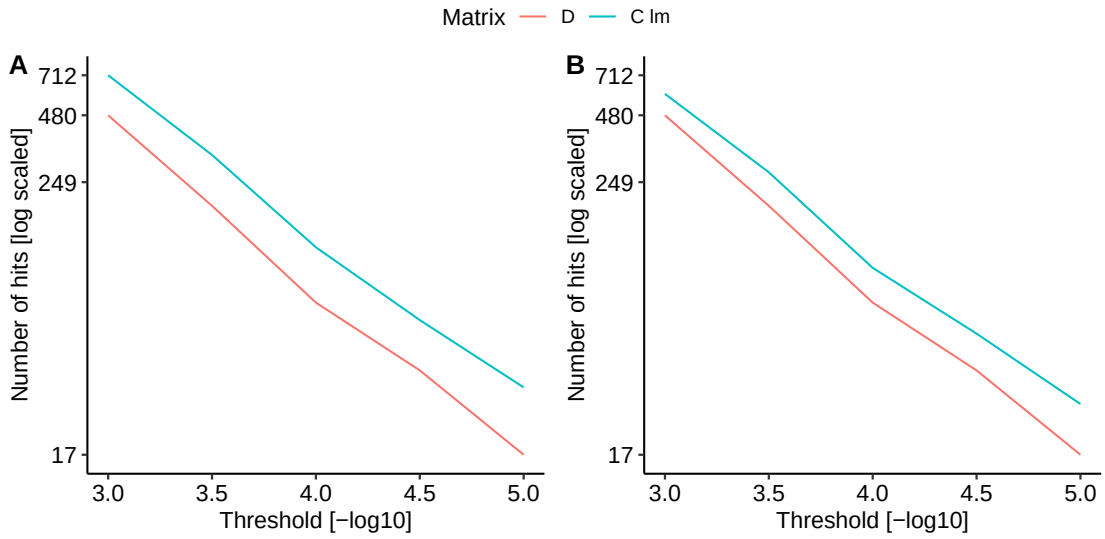

Figure 4.7: Number of significant hits in association tests for varying values of threshold. Different lines are plotted for the number of hits when using the *discrete* coding (matrix $\boldsymbol{D}$) and *continuous* coding using LM model. In **A** results from matrices $\boldsymbol{C}$ and $\boldsymbol{C_2}$ are combined. In **B** values from $\boldsymbol{C_2}$ are not included in the *continuous* coding results.

trait in which the test was significant is shown, Fiber from first harvest late. With the GAM modeling there was an increase of between 6% and 16% in the number of hits found, in relation to the number of significant hits using the LM model (table 4.2).

Table 4.2: Number of significant tests using GAM model that were not significant in LM model, for a range of thresholds. Only results for *continuous* coding are considered.

Threshold $[-log_{10}]$	# hits	% increase in hits in relation to LM
3.0	111	15.6

Threshold [$-log_{10}$]	# hits	% increase in hits in relation to LM
3.5	31	9.5
4.0	14	10.7
4.5	5	7.8
5.0	2	6.1

From the results in figs. 4.5 and 4.6, fiber measured at 2nd Harvest and late season showed larger effect associations, for both matrices $\boldsymbol{D}$ and $\boldsymbol{R}$. In fig. 4.6 the most significant associations came from markers that mostly aligned to one of the arms of chromosome 9, and both matrices showed significant markers at this peak.

4.3.3 Genomic Prediction

The prediction accuracy results are presented in fig. 4.9, for all combinations of fiber traits, prediction methods, and the matrices $\boldsymbol{D}$, $\boldsymbol{C}$ and $\boldsymbol{C_2}$. Using matrix $\boldsymbol{D}$, prediction accuracy values varied from 0.29 to 0.37, for matrix $\boldsymbol{C}$ values varied from 0.3 to 0.41, and for matrix $\boldsymbol{C_2}$ values varied from 0.29 to 0.41. With regard to the prediction methods, table 4.3 shows the method that had the best performance for each combination of trait and molecular marker matrix. In half the cases, GBLUP was the best one, and RF was never the best. For most cases, excluding RF, the difference is small between the methods. Table 4.4 shows the best matrix for all combinations of traits and methods. In all but two cases, the markers analyzed as *continuous* coding (matrices $\boldsymbol{C}$ and $\boldsymbol{C_2}$) had the best results. It is noticeable that, despite the smaller number of molecular markers used in

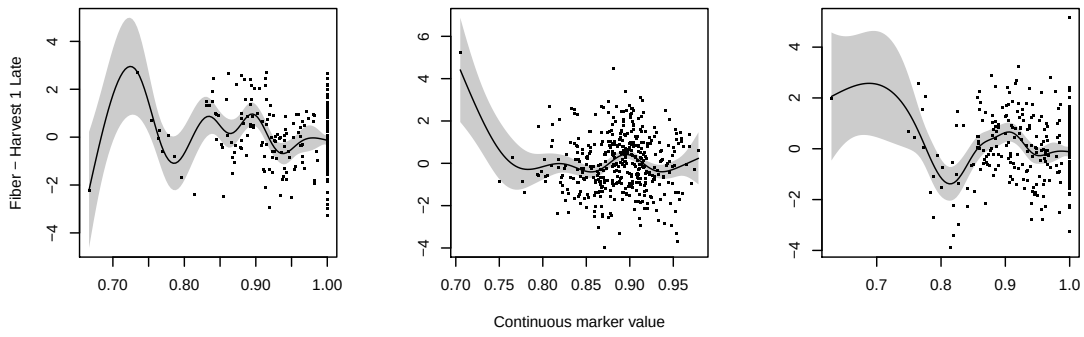

Figure 4.8: Trait values of Fiber (Harvest 1 late) plotted against marker values in continuous coding. Three examples of markers that were significant (p-value < 10^{-4}) for GAM model, but were not significant for LM model (p-value < 0.05) are shown. Continuous lines show the estimates for the smooth functions fitted in the GAM model, and the shadowed regions delimit confidence bands at two standard deviation above and bellow the estimate of the smooth functions.

matrix $\boldsymbol{C_2}$, for GBLUP, RKHS and SVM, its prediction accuracies were better than using matrix $\boldsymbol{D}$ for both Early season cases (fig. 4.9, panels on the left), and not much smaller in the Late cases (fig. 4.9, panels on the right).

Given the small differences between the best performing prediction methods, and the most common best method being GBLUP, we analyzed the prediction accuracy in a multiple-kernel setting focused on the GBLUP method. Here, we included either the matrix $\boldsymbol{C}$ or the matrix $\boldsymbol{C_2}$ as a second random predictor term (kernel) in the GBLUP model. Figure 4.10 shows a comparison of the predictions using multiple-kernel GBLUP and the regular GBLUP (results already shown in fig. 4.9). The use of the second kernel did not improve significantly in relation to the best performing results with single kernel.

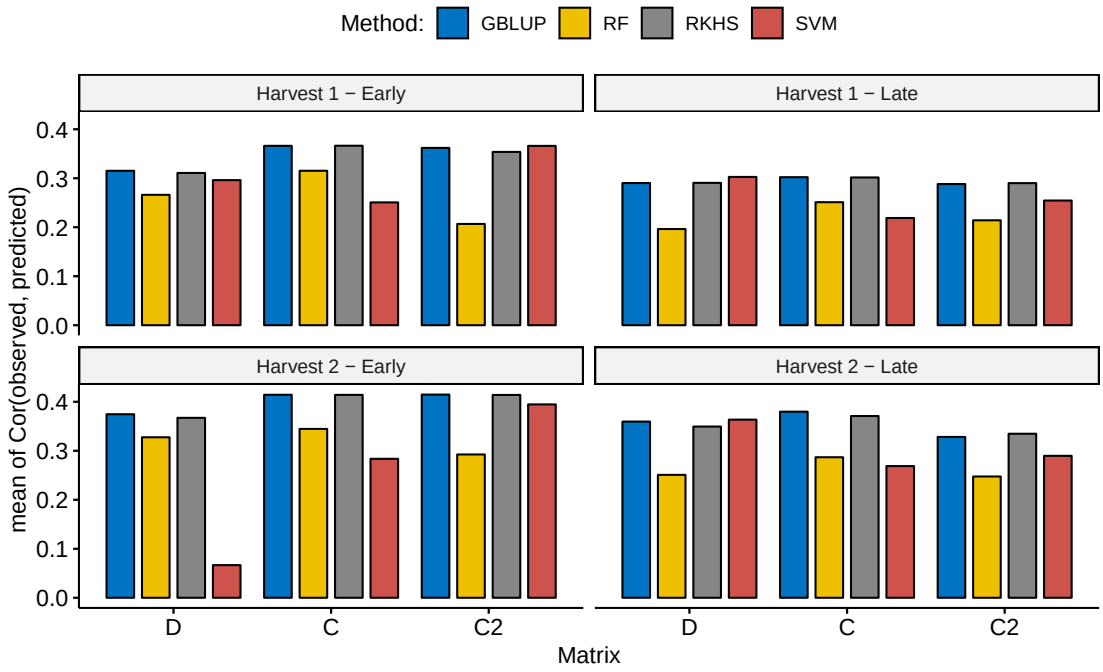

Figure 4.9: Mean prediction accuracy in cross validation with eight folds and five replications, for the trait Fiber under four conditions. Predictions were based on markers under *discrete* coding (matrix $\boldsymbol{D}$) or *continuous* coding (matrices $\boldsymbol{C}$ and $\boldsymbol{C_2}$), and for each coding, the methods GBLUP, RKHS, SVM and RF were used.

Table 4.3: Method that provided the best prediction result for each combination of Trait and Matrix.

trait	$\boldsymbol{D}$	$\boldsymbol{C}$	$\boldsymbol{C_2}$
Harvest 1 - Early	GBLUP	RKHS	SVM
Harvest 1 - Late	SVM	GBLUP	RKHS
Harvest 2 - Early	GBLUP	GBLUP	GBLUP
Harvest 2 - Late	SVM	GBLUP	RKHS

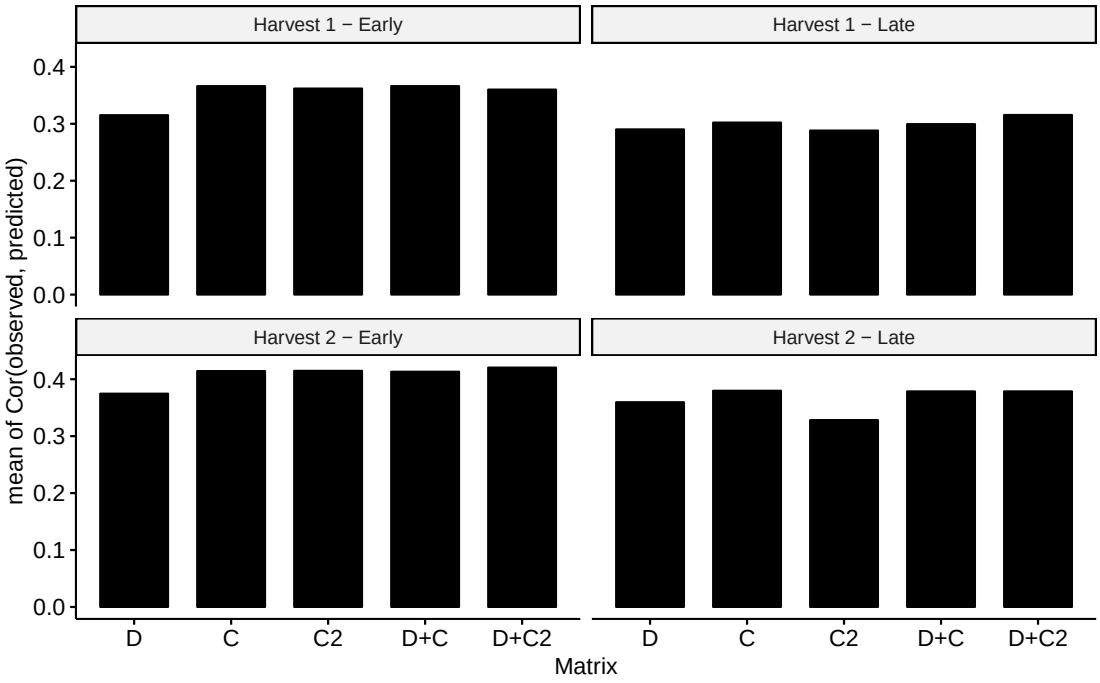

Figure 4.10: Mean prediction accuracy in Cross Validation. Predictions were based on model with single kernel (using matrices $\boldsymbol{D}$, $\boldsymbol{C}$ and $\boldsymbol{C_2}$) or two kernels (using $\boldsymbol{D}$+$\boldsymbol{C}$ or using $\boldsymbol{D}$+$\boldsymbol{C_2}$).

Table 4.4: Matrix that provided the best prediction result for each combination of Trait and Method.

trait	GBLUP	RF	RKHS	SVM
Harvest 1 - Early	$\boldsymbol{C}$	$\boldsymbol{C}$	$\boldsymbol{C}$	$\boldsymbol{C_2}$
Harvest 1 - Late	$\boldsymbol{C}$	$\boldsymbol{C}$	$\boldsymbol{C}$	$\boldsymbol{D}$
Harvest 2 - Early	$\boldsymbol{C_2}$	$\boldsymbol{C}$	$\boldsymbol{C}$	$\boldsymbol{C_2}$
Harvest 2 - Late	$\boldsymbol{C}$	$\boldsymbol{C}$	$\boldsymbol{C}$	$\boldsymbol{D}$

4.4 Discussion

In this study we analyzed the effect of using a molecular marker coding that might use dosage information for alleles. Sugarcane was used to evaluate this, taking advantage of its high ploidy level and availability of molecular marker sequencing data. We used two coding systems for the molecular markers, one that makes a simplified representation of the marker data (*discrete* coding), and another that makes a representation that is potentially influenced by the dosage of observed alleles (*continuous* coding).

Our data do not allow the independent verification of how read counts correlate to actual allele dosage. This is an assumption in order to assert that ratios used in the *continuous* coding represent the dosage ratio between alleles, which is a claim that cannot be made in the current analysis. On the other hand, the reproducibility of ratios in the same genotype observed in comparison to random pairs of different genotypes (fig. 4.2) shows that the ratios may encode relevant information. The relationship between individuals was not changed significantly when observed in the PCA (fig. 4.4, matrix $\boldsymbol{C}$ in comparison to matrix $\boldsymbol{D}$). It remains to be tested whether a dataset with higher read depth would produce better results, but the current results provide evidence to support the design of experiments using higher read depth in sugarcane.

The use of coding systems that would classify the marker data in clusters associated to individual allele dosage, as exemplified by Cordeiro et al. (2006) and Garcia et al. (2013), may lead to cases where a portion of the molecular markers cannot be classified due to lack of fit to the models assumed. This might result in datasets with reduced numbers of markers. It might also be

required to manually review the classification results, hindering the use of such a method in high throughput settings. On the other hand, the use of a *continuous* coding, which is based on few assumptions without a modeling of the observed marker data, led to an increase in the number of polymorphic makers in the current project, which were separated to the $\boldsymbol{C_2}$ matrix. There was also evidence that the markers under *continuous* coding were able to encode information not present in the *discrete* coding, as revealed by the lower correlation between off-diagonal elements from the relationship matrices derived from the matrices $\boldsymbol{D}$ and $\boldsymbol{C_2}$.

Recently, Endelman et al. (2018) demonstrated the advantage of using dosage information in GP for the tetraploid potato. In sugarcane, Garcia et al. (2006) has demonstrated the use of markers having a 1:3 segregation ratio (which can be derived from double dosage alleles, or a single dose allele from both parents from a biparental cross) for linkage mapping, but tests in association with traits or genomic prediction was lacking. Here we observed an increase in the number of significant hits in association tests when using *dosage* coding. In GP, results showed a small difference in prediction accuracy in the comparison between *continuous* and *discrete* coding, but in most of the cases there was an advantage when using *continuous* coding. We observed cases when even a smaller dataset of *continuous* coded markers was able to outperform the larger *discrete* coded dataset (fig. 4.10). In general, there was advantage when using *continuous* coding in the analyses that related phenotypic traits to molecular marker information (GWAS and GP) when dosage information was attempted to be included through the use of *continuous* coding.

The use of a test based on non-linear models (GAM) allowed to find a few

more hits over the ones already found found using a linear model, in GWAS, demonstrating how the *continuous* coding can take advantage of non-linear regression modeling. For GP, the use of models that would explore non-linear relationships between the predictors (SVM, RF and RKHS) did not outperform the GBLUP model. The traits used for this study were previously observed to be quantitatively inherited and controlled by additive genetic effects (Hogarth, 1987), which could be a factor limiting the accuracy of the non-linear models.

The hypothesis of this study was supported in the analyses performed here, and so the use of molecular marker codings that are based on dosage information can be beneficial in quantitative trait analysis for sugarcane.

BIBLIOGRAPHY

K. S. Aitken, P. A. Jackson, and C. L. McIntyre. Quantitative trait loci identified for sugar related traits in a sugarcane (Saccharum spp.) cultivar x Saccharum officinarum population. *Theoretical and Applied Genetics*, 112(7):1306–1317, 2006. ISSN 00405752. doi: 10.1007/s00122-006-0233-2.

K S Aitken, P a Jackson, and C L McIntyre. Construction of a genetic linkage map for Saccharum officinarum incorporating both simplex and duplex markers to increase genome coverage. *Genome*, 50(8):742–756, 2007. ISSN 0831-2796. doi: 10.1139/g07-056.

K S Aitken, S Hermann, K Karno, G D Bonnett, L. C. McIntyre, and P A Jackson. Genetic control of yield related stalk traits in sugarcane. *Theoretical and Applied Genetics*, 117(7):1191–1203, 2008. ISSN 00405752. doi: 10.1007/s00122-008-0856-6.

S. M. Aljanabi, Y. Parmessur, H. Kross, S. Dhayan, S. Saumtally, K. Ramdoyal, L. J C Autrey, and A. Dookun-Saumtally. Identification of a major quantitative trait locus (QTL) for yellow spot (Mycovellosiella koepkei) disease resistance in sugarcane. *Molecular Breeding*, 19(1):1–14, 2007. ISSN 13803743. doi: 10.1007/s11032-006-9008-3.

Felicity Claire Atkin, Mark J. Dieters, and Joanne K Stringer. Impact of depth of pedigree and inclusion of historical data on the estimation of additive variance and breeding values in a sugarcane breeding program. *Theoretical and Applied Genetics*, 119(3):555–565, 2009. ISSN 00405752. doi: 10.1007/s00122-009-1065-7. URL `https://link.springer.com/article/10.1007{%}2Fs00122-009-1065-7`.

Peter Baker, Phillip Jackson, and Karen Aitken. Bayesian estimation of marker dosage in sugarcane and other autopolyploids. *Theoretical and Applied Genetics*, 120(8):1653–1672, 2010. ISSN 00405752. doi: 10.1007/s00122-010-1283-z.

Nandita Banerjee, Archana Siraree, Sonia Yadav, Sanjeev Kumar, J. Singh, Sanjeev Kumar, Dinesh K. Pandey, and Ram K. Singh. Marker-trait association study for sucrose and yield contributing traits in sugarcane (Saccharum spp. hybrid). *Euphytica*, 205(1):185–201, 2015. ISSN 15735060. doi: 10.1007/s10681-015-1422-3. URL http://dx.doi.org/10.1007/s10681-015-1422-3.

J. A. Bressiani, R. Vencovsky, and W. L. Burnquist. Modified Sequential Selection in Sugarcane. In *International Society of Sucar Cane Technologists*, volume 25, pages 459–467, 2005. URL http://www.issct.org/pdf/proceedings/2005/2005BressianiBreedingSugarcaneforLeafScaldResistance.pdf.

A. H D Brown, J. Daniels, and B. D H Latter. Quantitative genetics of sugarcane - II. Correlation analysis of continuous characters in relation to hybrid sugarcane breeding. *Theoretical and Applied Genetics*, 39(1):1–10, 1969. ISSN 00405752. doi: 10.1007/BF00283078.

Itaraju Junior Baracuhy Brum, Karine Miranda Oliveira, Francisco Claudio Lopes, Thiago Romanos Benatti, and Mark E. Sorrells. Performance of Genomic Prediction for a Sugarcane Commercial Breeding Program. *(In preparation)*, 2018.

Peter C. Bundock, Rosanne E. Casu, and Robert J. Henry. Enrichment of genomic DNA for polymorphism detection in a non-model highly polyploid

crop plant. *Plant Biotechnology Journal*, 10(6):657–667, 2012. ISSN 14677644. doi: 10.1111/j.1467-7652.2012.00707.x.

M. P L Calus and R. F. Veerkamp. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics*, 124(6):362–368, 2007. ISSN 09312668. doi: 10.1111/j.1439-0388.2007.00691.x.

Giovanni M Cordeiro, Frances Eliott, C. Lynne McIntyre, Rosanne E Casu, and Robert J Henry. Characterisation of single nucleotide polymorphisms in sugarcane ESTs. *Theoretical and Applied Genetics*, 113(2):331–343, 2006. ISSN 00405752. doi: 10.1007/s00122-006-0300-8.

L. Costet, L. Le Cunff, S. Royaert, L. M. Raboin, C. Hervouet, L. Toubi, H. Telismart, O. Garsmeur, Y. Rousselle, J. Pauquet, S. Nibouche, J. C. Glaszmann, J. Y. Hoarau, and A. D'Hont. Haplotype structure around Bru1 reveals a narrow genetic basis for brown rust resistance in modern sugarcane cultivars. *Theoretical and Applied Genetics*, 125(5):825–836, 2012. ISSN 00405752. doi: 10.1007/s00122-012-1875-x.

José Crossa, Gustavo de Los Campos, Paulino Pérez, Daniel Gianola, Juan Burgueño, José Luis Araus, Dan Makumbi, Ravi P Singh, Susanne Dreisigacker, Jianbing Yan, Vivi Arief, Marianne Banziger, and Hans-Joachim Braun. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, 186(2):713–24, Oct 2010. doi: 10.1534/genetics.110.118521.

Julie C. Dawson, Jeffrey B. Endelman, Nicolas Heslot, Jose Crossa, Jesse Poland, Susanne Dreisigacker, Yann Manès, Mark E. Sorrells, and Jean Luc Jannink.

The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Research*, 154:12–22, 2013. ISSN 03784290. doi: 10.1016/j.fcr.2013.07.020. URL http://dx.doi.org/10.1016/j.fcr.2013.07.020.

Angélique D'Hont, Florence Paulet, and Jean Christophe Glaszmann. Oligoclonal interspecific origin of 'north indian' and 'chinese' sugarcanes. *Chromosome Research*, 10(3):253–262, Mar 2002. ISSN 1573-6849. doi: 10.1023/A:1015204424287. URL https://doi.org/10.1023/A:1015204424287.

J. B. Endelman and J.-L. Jannink. Shrinkage Estimation of the Realized Relationship Matrix. *G3: Genes|Genomes|Genetics*, 2(11):1405–1413, 2012. ISSN 2160-1836. doi: 10.1534/g3.112.004259. URL http://g3journal.org/cgi/doi/10.1534/g3.112.004259.

Jeffrey B. Endelman, Cari A. Schmitz Carley, Paul C. Bethke, Joseph J. Coombs, Mark E. Clough, Washington L. da Silva, Walter S. De Jong, David S. Douches, Curtis M. Frederick, Kathleen G. Haynes, and et al. Genetic variance partitioning and genome-wide prediction with allele dosage information in autotetraploid potato. *Genetics*, 209(1):77–87, Mar 2018. ISSN 1943-2631. doi: 10.1534/genetics.118.300685. URL http://dx.doi.org/10.1534/genetics.118.300685.

FAO. Faostat. production (crops). (latest update: 28-june-2018), 2018. URL http://www.fao.org/faostat/en/#data/QC.

FAOSTAT. Food and Agricultural commodities production - Commodities by country - World., 2016. URL http://faostat3.fao.org/browse/rankings/commodities{_}by{_}regions/.

A. A F Garcia, E. A. Kido, A. N. Meza, H. M B Souza, L. R. Pinto, M. M. Pastina, C. S. Leite, J. A G Da Silva, E. C. Ulian, A. Figueira, and A. P. Souza. Development of an integrated genetic map of a sugarcane (Saccharum spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. *Theoretical and Applied Genetics*, 112(2):298–314, 2006. ISSN 00405752. doi: 10.1007/s00122-005-0129-6.

Antonio A.F. Garcia, Marcelo Mollinari, Thiago G. Marconi, Oliver R. Serang, Renato R. Silva, Maria L.C. Vieira, Renato Vicentini, Estela A. Costa, Melina C. Mancini, Melissa O.S. Garcia, Maria M. Pastina, Rodrigo Gazaffi, Eliana R.F. Martins, Nair Dahmer, Danilo A. Sforça, Claudio B.C. Silva, Peter Bundock, Robert J. Henry, Glaucia M. Souza, Marie Anne Van Sluys, Marcos G.A. Landell, Monalisa S. Carneiro, Michel A.G. Vincentz, Luciana R. Pinto, Roland Vencovsky, and Anete P. Souza. SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Scientific Reports*, 3:1–10, 2013. ISSN 20452322. doi: 10.1038/srep03399.

Olivier Garsmeur, Gaetan Droc, Rudie Antonise, Jane Grimwood, Bernard Potier, Karen Aitken, Jerry Jenkins, Guillaume Martin, Carine Charron, Catherine Hervouet, and et al. A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nature Communications*, 9(1), Jul 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05051-5. URL `http://dx.doi.org/10.1038/s41467-018-05051-5`.

R. Chris Gaynor, Gregor Gorjanc, Alison R. Bentley, Eric S. Ober, Phil Howell, Robert Jackson, Ian J. Mackay, and John M. Hickey. A two-part strategy for using genomic selection to develop inbred lines. *Crop Science*, 57(5):2372, 2017.

ISSN 0011-183X. doi: 10.2135/cropsci2016.09.0742. URL http://dx.doi.org/10.2135/cropsci2016.09.0742.

Arthur R Gilmour, Robin Thompson, and Brian R Cullis. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics*, 51(4):1440–1450, 1995.

José Goldemberg, Suani Teixeira Coelho, and Patricia Guardabassi. The sustainability of ethanol production from sugarcane. *Energy Policy*, 36 (6):2086–2097, 2008. ISSN 03014215. doi: 10.1016/j.enpol.2008.02.028. URL http://www.sciencedirect.com/science/article/pii/S0301421508001080?via{%}3Dihub.

M. Gouy, Y. Rousselle, D. Bastianelli, P. Lecomte, L. Bonnal, D. Roques, J. C. Efile, S. Rocher, J. Daugrois, L. Toubi, S. Nabeneza, C. Hervouet, H. Telismart, M. Denis, A. Thong-Chane, J. C. Glaszmann, J. Y. Hoarau, S. Nibouche, and L. Costet. Experimental assessment of the accuracy of genomic selection in sugarcane. *Theoretical and Applied Genetics*, 126:2575–2586, 2013. ISSN 00405752. doi: 10.1007/s00122-013-2156-z.

M. Gouy, Y. Rousselle, A. Thong Chane, A. Anglade, S. Royaert, S. Nibouche, and L. Costet. Genome wide association mapping of agro-morphological and disease resistance traits in sugarcane. *Euphytica*, 202(2):269–284, 2015. ISSN 15735060. doi: 10.1007/s10681-014-1294-y.

Laurent Grivet, Angélique D'Hont, Danièle Roques, Philippe Feldmann, Claire Lanaud, and Jean Christophe Glaszmann. RFLP mapping in cultivated sugarcane (Saccharum spp.): Genome organization in a highly polyploid and aneuploid interspecific hybrid. *Genetics*, 142(3):987–1000, 1996. ISSN 00166731.

D. Habier, R. L. Fernando, and J. C M Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4): 2389–2397, 2007. ISSN 00166731. doi: 10.1534/genetics.107.081190.

Nicolas Heslot, Deniz Akdemir, Mark E. Sorrells, and Jean Luc Jannink. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics*, 127(2):463–480, 2014. ISSN 00405752. doi: 10.1007/s00122-013-2231-5.

Nicolas Heslot, Jean-Luc Jannink, and Mark E. Sorrells. Perspectives for Genomic Selection Applications and Research in Plants. *Crop Science*, 55(february):1–12, 2015. ISSN 0011-183X. doi: 10.2135/cropsci2014.03.0249. URL `https://dl.sciencesocieties.org/publications/cs/abstracts/55/1/1`.

Nam V. Hoang, Agnelo Furtado, Frederik C. Botha, Blake A. Simmons, and Robert J. Henry. Potential for Genetic Improvement of Sugarcane as a Source of Biomass for Biofuels. *Frontiers in Bioengineering and Biotechnology*, 3(November):1–15, 2015. ISSN 2296-4185. doi: 10.3389/fbioe.2015.00182. URL `http://journal.frontiersin.org/Article/10.3389/fbioe.2015.00182/abstract`.

J. Y. Hoarau, L. Grivet, B. Offmann, L. M. Raboin, J. P. Diorflar, J. Payet, M. Hellmann, A. D'Hont, and J. C. Glaszmann. Genetic dissection of a modern sugarcane cultivar (Saccharum spp.). II. Detection of QTLs for yield components. *Theoretical and Applied Genetics*, 105(6-7):1027–1037, 2002. ISSN 00405752. doi: 10.1007/s00122-002-1047-5.

D M Hogarth. Genetics of Sugarcane. In Don J Heinz, editor, *Sugar-*

cane Improvement through Breeding, volume 11 of *Developments in Crop Science*, chapter 6, pages 255–271. Elsevier, 1987. doi: https://doi.org/10.1016/B978-0-444-42769-4.50011-4. URL http://www.sciencedirect.com/science/article/pii/B9780444427694500114.

J. E. Irvine. Saccharum species as horticultural classes. *Theoretical and Applied Genetics*, 98(2):186–194, 1999. ISSN 00405752. doi: 10.1007/s001220051057.

Phillip A Jackson. Breeding for improved sugar content in sugarcane. *Field Crops Research*, 92:277–290, 2005. doi: 10.1016/j.fcr.2005.01.024.

Diego Jarquín, José Crossa, Xavier Lacaze, Philippe Du Cheyron, Joëlle Daucourt, Josiane Lorgeou, François Piraux, Laurent Guerreiro, Paulino Pérez, Mario Calus, Juan Burgueño, and Gustavo de los Campos. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, 127(3):595–607, 2014. ISSN 00405752. doi: 10.1007/s00122-013-2243-1.

Philomin Juliana, Ravi P Singh, Pawan K Singh, Jose Crossa, Julio Huerta-Espino, Caixia Lan, Sridhar Bhavani, Jessica E Rutkoski, Jesse A Poland, Gary C Bergstrom, and Mark E Sorrells. Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat. *Theor Appl Genet*, 130(7):1415–1430, Jul 2017. doi: 10.1007/s00122-017-2897-1.

C a Kimbeng and M C Cox. Early Generation Selection of Sugarcane Families and Clones in Australia: a Review. *Journal american society of sugarcane technologist*, 23:20–39, 2003.

R. Lande and R. Thompson. Efficiency of marker-assisted selection in the

improvement of quantitative traits. *Genetics*, 124(3):743–756, 1990. ISSN 00166731.

Loïc Le Cunff, Olivier Garsmeur, Louis Marie Raboin, Jérome Pauquet, Hugues Telismart, Athiappan Selvi, Laurent Grivet, Romain Philippe, Dilara Begum, Monique Deu, Laurent Costet, Rod Wing, Jean Christophe Glaszmann, and Angélique D'Hont. Diploid/polyploid syntenic shuttle mapping and haplotype-specific chromosome walking toward a rust resistance gene (Bru1) in highly polyploid sugarcane (2n ˜ 12x ˜ 115). *Genetics*, 180(1):649–660, 2008. ISSN 00166731. doi: 10.1534/genetics.108.091355.

A Legarra, I Aguilar, and I Misztal. A relationship matrix including full pedigree and genomic information. *Journal of dairy science*, 92(9):4656–4663, 2009. ISSN 1525-3198. doi: 10.3168/jds.2009-2061. URL `http://dx.doi.org/10.3168/jds.2009-2061`.

Pingwu Liu, Amaresh Chandra, Youxiong Que, Ping Hua Chen, Michael P. Grisham, William H. White, Caleb D. Dalley, Thomas L. Tew, and Yong Bao Pan. Identification of quantitative trait loci controlling sucrose content based on an enriched genetic linkage map of sugarcane (Saccharum spp. hybrids) cultivar 'LCP 85-384'. *Euphytica*, 207(3):527–549, 2016. ISSN 15735060. doi: 10.1007/s10681-015-1538-5.

Aaron J. Lorenz, Shiaoman Chao, Franco G. Asoro, Elliot L. Heffner, Takeshi Hayashi, Hiroyoshi Iwata, Kevin P. Smith, Mark E. Sorrells, and Jean-Luc Jannink. Genomic selection in plant breeding: Knowledge and prospects. In Donald L. Sparks, editor, *Advances in Agronomy*, volume 110 of *Advances in Agronomy*, pages 77–123. Academic Press, 2011. ISBN 9780123855312. doi: 10.1016/B978-0-12-385531-2.00002-5.

C. L. McIntyre, V. A. Whan, B. Croft, R. Magarey, and G. R. Smith. Identification and validation of molecular markers associated with Pachymetra root rot and brown rust resistance in sugarcane using map- and association-based approaches. *Molecular Breeding*, 16(2):151–161, 2005. ISSN 13803743. doi: 10.1007/s11032-005-7492-5.

T H E Meuwissen, B J Hayes, and M E Goddard. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157:1819–1829, 2001. ISSN 0016-6731. doi: 11290733.

Karin Meyer. WOMBAT - A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *J Zhejiang Univ Sci B*, 8(11):815–821, 2007. ISSN 1673-1581. doi: 10.1631/jzus.2007.B0815. URL 10.1631/jzus.2007.B0815{%}5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC2064953/pdf/JZUSB08-0815.pdf.

S. B. Milligan, F. A. Martin, and K. A. Gravois. Inheritance of sugarcane ratooning ability and the relationship of younger crop traits to older crop traits. *Crop Science*, 36(1):45, 1996. ISSN 0011-183X. doi: 10.2135/cropsci1996.0011183x003600010008x. URL http://dx.doi.org/10.2135/cropsci1996.0011183X003600010008x.

R Ming, S. C. Liu, Y. R. Lin, J. Da Silva, W Wilson, D Braga, A. Van Deynze, T F Wenslaff, K K Wu, P H Moore, W Burnquist, M E Sorrells, J E Irvine, and A H Paterson. Detailed alignment of Saccharum and Sorghum chromosomes: Comparative organization of closely related diploid and polyploid genomes. *Genetics*, 150(4):1663–1682, 1998. ISSN 00166731. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1460436/.

R. Ming, Y. W. Wang, X. Draye, P. H. Moore, J. E. Irvine, and A. H. Pater-

son. Molecular dissection of complex traits in autopolyploids: Mapping QTLs affecting sugar yield and related traits in sugarcane. *Theoretical and Applied Genetics*, 105(2-3):332–345, 2002a. ISSN 00405752. doi: 10.1007/s00122-001-0861-5.

Ray Ming, Sin Chieh Liu, Paul H. Moore, James E. Irvine, and Andrew H. Paterson. QTL analysis in a complex autopolyploid: Genetic control of sugar content in sugarcane. *Genome Research*, 11(12):2075–2084, 2001. ISSN 10889051. doi: 10.1101/gr.198801.

Ray Ming, Terrye A Del Monte, Eduardo Hernandez, Paul H Moore, James E Irvine, and Andrew H Paterson. Comparative analysis of QTLs affecting plant height and flowering among closely-related diploid and polyploid genomes. *Genome*, 45(5):794–803, 2002b. ISSN 0831-2796. doi: 10.1139/g02-042.

Ray Ming, Paul H. Moore, Kuo Kao Wu, Angélique D'Hont, Jean C. Glaszmann, Thomas L. Tew, T. Erik Mirkov, Jorge da Silva, John Jifon, Mamta Rai, Raymond J. Schnell, Stevens M. Brumbley, Prakash Lakshmanan, Jack C. Comstock, and Andrew H. Paterson. Sugarcane Improvement through Breeding and Biotechnology. In Jules Janick, editor, *Plant Breeding Reviews*, volume 27, chapter 2, pages 15–118. Wiley-Blackwell, 2010. ISBN 9780470650349. doi: 10.1002/9780470650349.ch2. URL `https://onlinelibrary.wiley.com/doi/10.1002/9780470650349.ch2https://doi.org/10.1002/9780470650349.ch2`.

Patricio R. Munoz, Marcio F.R. Resende, Dudley A. Huber, Tania Quesada, Marcos D.V. Resende, David B. Neale, Jill L. Wegrzyn, Matias Kirst, and Gary F. Peter. Genomic relationship matrix for correcting pedigree errors in breeding populations: Impact on genetic parameters and genomic selec-

tion accuracy. *Crop Science*, 54(3):1115–1123, 2014. ISSN 14350653. doi: 10.2135/cropsci2012.12.0673.

Vagner Katsumi Okura, Rafael S. C. de Souza, Susely F de Siqueira Tada, and Paulo Arruda. BAC-Pool Sequencing and Assembly of 19 Mb of the Complex Sugarcane Genome. *Frontiers in Plant Science*, 7, 2016. ISSN 1664-462X. doi: 10.3389/fpls.2016.00342. URL http://journal.frontiersin.org/Article/10.3389/fpls.2016.00342/abstract.

M M Pastina, M Malosetti, R Gazaffi, M Mollinari, G R A Margarido, K M Oliveira, L R Pinto, A P Souza, F A van Eeuwijk, and A A F Garcia. A mixed model QTL analysis for sugarcane multiple-harvest-location trial data. *Theoretical and Applied Genetics*, 124(5):835–849, mar 2012. ISSN 1432-2242. doi: 10.1007/s00122-011-1748-8. URL https://doi.org/10.1007/s00122-011-1748-8.

Andrew H Paterson, John E Bowers, Rémy Bruggmann, Inna Dubchak, Jane Grimwood, Heidrun Gundlach, Georg Haberer, Uffe Hellsten, Therese Mitros, Alexander Poliakov, Jeremy Schmutz, Manuel Spannagl, Haibao Tang, Xiyin Wang, Thomas Wicker, Arvind K Bharti, Jarrod Chapman, F Alex Feltus, Udo Gowik, Igor V Grigoriev, Eric Lyons, Christopher a Maher, Mihaela Martis, Apurva Narechania, Robert P Otillar, Bryan W Penning, Asaf a Salamov, Yu Wang, Lifang Zhang, Nicholas C Carpita, Michael Freeling, Alan R Gingle, C Thomas Hash, Beat Keller, Patricia Klein, Stephen Kresovich, Maureen C McCann, Ray Ming, Daniel G Peterson, Mehboob ur Rahman, Doreen Ware, Peter Westhoff, Klaus F X Mayer, Joachim Messing, and Daniel S Rokhsar. The Sorghum bicolor genome and the diversification of grasses. *Nature*, 457(7229):551–556, 2009. ISSN 0028-0836. doi:

10.1038/nature07723. URL http://www.ncbi.nlm.nih.gov/pubmed/19189423.

Nick Patterson, Alkes L Price, and David Reich. Population Structure and Eigenanalysis. *PLOS Genetics*, 2(12):1–20, 2006. doi: 10.1371/journal.pgen.0020190. URL https://doi.org/10.1371/journal.pgen.0020190.

H. P. Piepho, J. Möhring, A. E. Melchinger, and A. Büchse. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161(1-2):209–228, 2008. ISSN 00142336. doi: 10.1007/s10681-007-9449-8.

George Piperidis, Nathalie Piperidis, and Angélique D'Hont. Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. *Molecular Genetics and Genomics*, 284(1):65–73, 2010. ISSN 16174615. doi: 10.1007/s00438-010-0546-3.

Nathalie Piperidis, Phillip A. Jackson, Angelique D'Hont, Pascale Besse, Jean Yves Hoarau, Brigitte Courtois, Karen S. Aitken, and C. Lynne McIntyre. Comparative genetics in sugarcane enables structured map enhancement and validation of marker-trait associations. *Molecular Breeding*, 21(2):233–247, 2008. ISSN 13803743. doi: 10.1007/s11032-007-9124-8.

R Core Team. R: A Language and Environment for Statistical Computing, 2016. URL https://www.r-project.org.

L. M. Raboin, K. M. Oliveira, L. Lecunff, H. Telismart, D. Roques, M. Butterfield, J. Y. Hoarau, and A. D'Hont. Genetic mapping in sugarcane, a high polyploid, using bi-parental progeny: Identification of a gene controlling stalk colour and a new rust resistance gene. *Theoretical and Applied Genetics*, 112(7):1382–1391, 2006. ISSN 00405752. doi: 10.1007/s00122-006-0240-3.

Josefina Racedo, Lucía Gutiérrez, María Francisca Perera, Santiago Ostengo, Esteban Mariano Pardo, María Inés Cuenya, Bjorn Welin, and Atilio Pedro Castagnaro. Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. *BMC Plant Biology*, 16(1):142, 2016. ISSN 1471-2229. doi: 10.1186/s12870-016-0829-x. URL http://bmcplantbiol.biomedcentral.com/articles/10.1186/s12870-016-0829-x.

Marcos D. V. Resende, Márcio F. R. Resende, Carolina P. Sansaloni, Cesar D. Petroli, Alexandre A. Missiaggia, Aurelio M. Aguiar, Jupiter M. Abad, Elizabete K. Takahashi, Antonio M. Rosado, Danielle A. Faria, Georgios J. Pappas, Andrzej Kilian, and Dario Grattapaglia. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytologist*, 194(1):116–128, apr 2012. ISSN 0028646X. doi: 10.1111/j.1469-8137.2011.04038.x. URL http://doi.wiley.com/10.1111/j.1469-8137.2011.04038.x.

Christian Riedelsheimer, Jeffrey B Endelman, Michael Stange, Mark E Sorrells, Jean-Luc Jannink, and Albrecht E Melchinger. Genomic Predictability of Interconnected Biparental Maize Populations. *Genetics*, 194(2):493 LP – 503, jun 2013. URL http://www.genetics.org/content/194/2/493.abstract.

M. I. Ripol, G. A. Churchill, J. A G Da Silva, and M Sorrells. Statistical aspects of genetic mapping in autopolyploids. *Gene*, 235(1-2):31–41, 1999. ISSN 03781119. doi: 10.1016/S0378-1119(99)00218-8.

G. K. Robinson. That BLUP is a Good Thing : The Estimation of Random Effects. *Statistical Science*, 6(1):15–32, 1991. ISSN 0883-4237. URL http://www.jstor.org/stable/2245695.

Jessica E Rutkoski, Jesse Poland, Jean-Luc Jannink, and Mark E Sorrells. Imputation of Unordered Markers and the Impact on Genomic Selection Accuracy. *G3: Genes|Genomes|Genetics*, 3(3):427 LP – 439, mar 2013. URL http://www.g3journal.org/content/3/3/427.abstract.

J C Skinner, D M Hogarth, and K K Wu. Selection Methods, Criteria, and Indices. In Don J Heinz, editor, *Sugarcane Improvement through Breeding*, volume 11 of *Developments in Crop Science*, chapter 11, pages 409–453. Elsevier, 1987. doi: https://doi.org/10.1016/B978-0-444-42769-4.50016-3. URL http://www.sciencedirect.com/science/article/pii/B9780444427694500163.

Jian Song, Xiping Yang, Marcio F. R. Resende, Leandro G. Neves, James Todd, Jisen Zhang, Jack C. Comstock, and Jianping Wang. Natural Allelic Variations in Highly Polyploidy Saccharum Complex. *Frontiers in Plant Science*, 7(June):1–18, 2016. ISSN 1664-462X. doi: 10.3389/fpls.2016.00804. URL http://journal.frontiersin.org/Article/10.3389/fpls.2016.00804/abstract.

J K Stringer, T A McRae, and M C Cox. Best Linear Unbiased Prediction as a Method of Estimating Breedig Value in Sugarcane. In Wilson JR, Hogarth DM, Campbell JA, and Garside AL, editors, *Sugarcane: Research Towards Efficient and Sustainable Production*, pages 39–41, Brisbane, 1996. CSIRO Division of Tropical Crops and Pastures.

J K Stringer, M C Cox, F C Atkin, X Wei, and D M Hogarth. Family Selection Improves the Efficiency and Effectiveness of Selecting Original Seedlings and Parents. *Sugar Tech*, 13(1):36–41, 2011. ISSN 0974-

0740. doi: 10.1007/s12355-011-0073-5. URL https://doi.org/10.1007/s12355-011-0073-5.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001. doi: 10.1093/bioinformatics/17.6.520. URL http://dx.doi.org/10.1093/bioinformatics/17.6.520http://smi-web.stanford.edu/projects/helix/pubs/impute/.

A. I. Vazquez, D. M. Bates, G. J M Rosa, D. Gianola, and K. A. Weigel. Technical note: An R package for fitting generalized linear mixed models in animal breeding. *Journal of Animal Science*, 88(2):497–504, 2010. ISSN 00218812. doi: 10.2527/jas.2009-1952.

André L. Vettore, Felipe R. da Silva, Edson L. Kemper, Glaucia M. Souza, Aline M. da Silva, Maria Inês T Ferro, Flavio Henrique-Silva, Éder A. Giglioti, Manoel V F Lemos, Luiz L. Coutinho, Marina P. Nobrega, Helaine Carrer, Suzelei C. França, Maurício Bacci, Maria Helena S Goldman, Suely L. Gomes, Luiz R. Nunes, Luis E A Camargo, Walter J. Siqueira, Marie Anne Van Sluys, Otavio H. Thiemann, Eiko E. Kuramae, Roberto V. Santelli, Celso L. Marino, Maria L P N Targon, Jesus A. Ferro, Henrique C S Silveira, Danyelle C. Marini, Eliana G M Lemos, Claudia B. Monteiro-Vitorello, José H M Tambor, Dirce M. Carraro, Patrícia G. Roberto, Vanderlei G. Martins, Gustavo H. Goldman, Regina C. de Oliveira, Daniela Truffi, Carlos A. Colombo, Magdalena Rossi, Paula G. de Araujo, Susana A. Sculaccio, Aline Angella, Marleide M A Lima, Vicente E. de Rosa, Fábio Siviero, Virginia E. Coscrato, Marcos A. Machado, Laurent Grivet, Sonia M Z Di Mauro, Francisco G. Nobrega,

Carlos F M Menck, Marilia D V Braga, Guilherme P. Telles, Frank A A Cara, Guilherme Pedrosa, João Meidanis, and Paulo Arruda. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Research*, 13(12):2725–2735, 2003. ISSN 10889051. doi: 10.1101/gr.1532103.

Jianping Wang, Bruce Roe, Simone Macmil, Qingyi Yu, Jan E Murray, Haibao Tang, Cuixia Chen, Fares Najar, Graham Wiley, John Bowers, Marie-Anne Van Sluys, Daniel S Rokhsar, Matthew E Hudson, Stephen P Moose, Andrew H Paterson, and Ray Ming. Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC genomics*, 11(1):261, 2010. ISSN 1471-2164. doi: 10.1186/1471-2164-11-261. URL `http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-261`.

Xianming Wei, Phillip A. Jackson, C. Lynne McIntyre, Karen S Aitken, and Barry Croft. Associations between DNA markers and resistance to diseases in sugarcane and effects of population substructure. *Theoretical and Applied Genetics*, 114(1):155–164, 2006. ISSN 00405752. doi: 10.1007/s00122-006-0418-8.

K. K. Wu, W. Burnquist, M. E. Sorrells, T. L. Tew, P. H. Moore, and S. D. Tanksley. The detection and estimation of linkage in polyploids using single-dose restriction fragments. *Theoretical and Applied Genetics*, 83(3):294–300, 1992. ISSN 00405752. doi: 10.1007/BF00224274.

Xiping Yang, Sushma Sood, Neil Glynn, Md. Sariful Islam, Jack Comstock, and Jianping Wang. Constructing high-density genetic maps for polyploid sugarcane (Saccharum spp.) and identifying quantitative trait loci controlling brown rust resistance. *Molecular Breeding*, 37(10):116, sep 2017. ISSN 1572-

9788. doi: 10.1007/s11032-017-0716-7. URL https://doi.org/10.1007/s11032-017-0716-7.

Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, Stephen Kresovich, and Edward S Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, feb 2006. ISSN 1061-4036. doi: 10.1038/ng1702. URL http://www.ncbi.nlm.nih.gov/pubmed/16380716http://www.nature.com/doifinder/10.1038/ng1702.

Jisen Zhang, Marvellous Zhou, James Walsh, Lin Zhu, Youqiang Chen, and Ray Ming. *Sugarcane Genetics and Genomics*, chapter 23, pages 623–643. Wiley-Blackwell, 2013. ISBN 9781118771280. doi: 10.1002/9781118771280.ch23. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118771280.ch23.

M. M. Zhou and M. L. Lichakane. Family selection gains for quality traits among South African sugarcane breeding populations. *South African Journal of Plant and Soil*, 29(3-4):143–149, 2012. ISSN 02571862. doi: 10.1080/02571862.2012.743606. URL https://doi.org/10.1080/02571862.2012.743606.

MM Zhou and MD Shoko. Simultaneous selection for yield and ratooning ability in sugarcane genotypes using analysis of covariance. *South African Journal of Plant and Soil*, 29(2):93–100, Sep 2012. ISSN 2167-034X. doi: 10.1080/02571862.2012.717639. URL http://dx.doi.org/10.1080/02571862.2012.717639.